

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

MMA: metadata supported multi-variate attention for onset detection and prediction

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/2037962> since 2024-12-13T15:43:20Z

Published version:

DOI:10.1007/s10618-024-01008-z

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

MMA: Metadata Supported Multi-variate Attention for Onset Detection and Prediction

Manjusha Ravindranath^{1*}, K. Selçuk Candan¹, Maria Luisa Sapino² and Brian Appavu³

¹*SCAI, Arizona State University, Tempe, Arizona, USA.

²Dipartimento di Informatica, University of Torino, Turin, Italy.

³*Neurology, Phoenix Children's, Phoenix, Arizona, USA.

*Corresponding author(s). E-mail(s): mravind1@asu.edu;
Contributing authors: candan@asu.edu; mlsapino@di.unito.it;
bappavu@phoenixchildrens.com;

Abstract

Deep learning has been applied successfully in sequence understanding and translation problems, especially in univariate, unimodal contexts, where large number of supervision data are available. The effectiveness of deep learning in more complex (multi-modal, multi-variate) contexts, where supervision data is rare, however, is generally not satisfactory. In this paper, we focus on improving detection and prediction accuracy in precisely such contexts – in particular, we focus on the problem of predicting seizure onsets relying on multi-modal (EEG, ICP, ECG, and ABP) sensory data streams, some of which (such as EEG) are inherently multi-variate due to the placement of multiple sensors to capture spatial distribution of the relevant signals. In particular, we note that multi-variate time series often carry robust, spatio-temporally localized features that could help predict onset events. We further argue that such features can be used to support implementation of metadata supported multivariate attention (or MMA) mechanisms that help significantly improve the effectiveness of neural networks architectures. In this paper, we use the proposed MMA approach to develop a multi-modal LSTM-based neural network architecture to tackle seizure onset detection and prediction tasks relying on EEG, ICP, ECG, and ABP data streams. We experimentally evaluate the proposed architecture under different scenarios – the results illustrate the effectiveness of the proposed attention mechanism, especially compared against other metadata driven competitors.

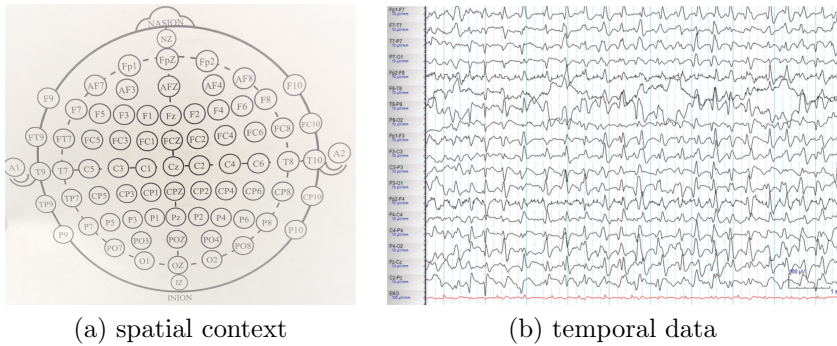


Fig. 1 Spatio-temporal encoding of EEG data for seizure early prediction – from the International 10-20 [3] (C=central, T=temporal, P=parietal, F=frontal, Fp=frontal polar, O=occipital)

Keywords: Multi-modal seizure onset prediction, rare event prediction, multi-variate attention

1 Introduction

Seizures are more wide-spread in population than most expects – about one percent of Americans have some form of epilepsy, and nearly four percent will develop epilepsy at some point in their lives [1]. Furthermore, the cumulative incidence of post-traumatic epilepsy (PTE) ranges widely, from 2% to over 50% depending on the severity of the injury [2].

In fact, seizures can be caused by a wide range of reasons and may occur due to diverse traumatic events (such as central nervous system infections, intracranial hemorrhage, stroke, brain injury, cancer or vitamin deficiencies [4–6]) and, consequently, can materialize in the form of diverse and unique spatio-temporal neurological patterns.

The most common types of seizures can be distinguished by the location where they begin within the brain. The seizures that involve networks in just one hemisphere of the brain are referred to as focal onset seizures while those that begin in both hemispheres of the brain are referred to as generalized onset seizures. If the events shift laterally they are called ping-pong seizures.

Seizure prediction requires modeling of complex non-linear spatio-temporal dynamics in various biological signals [7]. While evidence suggests that seizures are preceded by characteristic changes in the electroencephalogram (EEG) signals that are potentially detectable before the onset of a seizure [8], further evidence suggest that other sensory time series data, such as intracranial pressure (ICP), electrocardiogram (ECG), and arterial blood pressure (ABP) [9] are also very informative. Yet, despite the availability of multiple (multi-modal) data to help detect and predict onset of a seizure, the diversity and uniqueness of seizures pose a significant challenge. Deep learning has been applied

successfully in sequence understanding and translation problems, especially in univariate, unimodal contexts, where large number of supervision data are available [10]. The effectiveness of deep learning in more complex (multi-modal, multi-variate) contexts, such as seizure onset prediction, where supervision data is rare, however, is generally not satisfactory. Existing solutions fail to predict rare events despite recent advances, such as very rare seizure events in highly personalized post-traumatic EEG data as noted in [11]. Multi-variate time series like EEG carry robust localized multi-variate spatial features in addition to temporal features. These spatial features may help better identify these rare events when they have lateral patterns and are extremely rare like 0.5% at worst case. However, often there might not be enough data to train these events, so it is usually impossible to identify and use these features by neural architectures.

1.1 Our Contributions: Metadata supported Multivariate Attention (MMA)

In this paper, we focus on improving detection and prediction accuracy in precisely such contexts. In preliminary work, [11], we had proposed a *M2NN* model which extended the conventional single-layer LSTM architecture, with dual regional attention layers that performed context analysis across frequency channels for EEG data. Most existing work indeed relies on EEG data for seizure detection [7]. In contrast, we argue that a multi-modal approach to seizure forecasting is likely to produce more robust predictions and focus on the problem of predicting seizure onsets relying on multi-modal (EEG, ICP, ECG, and ABP) sensory data streams, some of which (such as EEG) are inherently multi-variate due to the placement of multiple sensors to capture the *spatial distribution* of the underlying signals (Figure 1). In particular, we note that multi-variate time series often carry robust, *spatio-temporally localized* features that could help predict onset events. We further argue that such features can be used to support implementation of metadata supported multivariate attention (or MMA) mechanisms that help significantly improve the effectiveness of neural networks architectures.

In this paper, we leverage the proposed MMA approach to develop a multi-modal LSTM-based neural network architecture to tackle seizure onset detection and prediction tasks relying on EEG, ICP, ECG, and ABP data streams. We can summarize the key contributions of our work as follows:

- We propose a multi-modal, multi-variate LSTM-based neural architecture that leverages a metadata supported multivariate attention (or MMA) mechanism that uses robust multi-variate spatio-temporal features that are extracted *a priori* - robust features are identified prior through a process external to the neural network architecture [12] and fed into the neural network as a side information. In particular, the robust multi-variate features [12] are extracted by simultaneously considering, at multiple scales, the temporal characteristics of the time series as well as external knowledge,

including variate relationships that are *a priori* known. MMA is supported with metadata describing the inter-relationships among the variates as explained in Section 2.1.1 to capture the contexts between the multiple variates from different sources to enable in-context learning. In this paper, we have considered two contexts namely *frequency context* and *frequency and spatial context*. In the second context, spatial context is explicitly considered when compared to the first where it is implicitly learned.

- We propose a deep learning approach which can leverage multi-variate features extracted from the EEG modality as an attention mechanism. These robust multivariate features are extracted outside of the neural network and fed into the network as side information. In our prior work [11], we had shown that such multi-variate features can be effectively extracted considering different frequency channels. In this paper, we further extend this approach to capture multi-variate features across multiple EEG sensors and experimentally show that the addition of spatial context leads to improved detection and prediction of rare seizure onsets. The proposed approach analyzes EEG data with *frequency context* and *frequency and spatial context* (with and without adaptive variate clustering) to predict the seizure onsets 5.1 minutes ahead of time. Adaptive variate clustering is proposed when compared to the fixed number of clusters used in [11]. This is another method intended mainly for patients with single and multiple seizure event clusters. The number of clusters are adaptively learnt per patient using silhouette score [13] with K-Means clustering; silhouette score being a way to measure how similar a data point is within a cluster compared to other clusters [13].
- The approach learns and interprets variates from multiple modalities like ICP, ECG and ABP (robust multivariate features are extracted considering only the frequency context as ICP, ECG and ABP have no spatial context), in addition to the EEG modality using a segmented LSTM (four separate LSTMs for each input data source or modality) model.
- In addition to direct learning on a particular patient, the proposed approach also is able to transfer learned models between patients - from a donor/provider patient to a test patient of similar or dissimilar category.

In summary, the main contribution of the paper is the rare event inference of seizure onsets. For this we have proposed an algorithm using metadata which is compared against other attentioned and non-attentioned baselines. Unlike our prior work [11], the approach considers multiple data modalities (ICP, ECG and ABP, in addition to EEG) and also takes into account both *frequency and spatial contexts* as metadata. We experimentally evaluate the proposed architecture under different scenarios described in detail in Section 3.1. Since our focus is seizure detection and prediction, the bulk of the experiments have focussed on seizure detection and they have shown that the proposed metadata-driven attention mechanism helps improve onset detection and prediction accuracies by helping to focus on the most informative segments of the multi-modal, multi-variate EEG, ICP, ECG, and ABP time series used in seizure detection and prediction. In particular, for the task of detecting the

preictal state (which appears before the seizure begins) five minutes before the onset of a seizure – the results, reported in Section 3.3.1, shows the effectiveness of the proposed attention mechanism, especially compared against other metadata driven competitors [14].

Unfortunately, the EEG, ICP, ECG, and ABP time series used for evaluation cannot be released due to HIPAA protections. We, therefore, included in the manuscript experiments with additional public data sets from other domains, including (a) COVID data, (b) traffic flow data, and (c) bitcoin price data, that share some of the common characteristics of the seizure onset detection data. In particular, both COVID and traffic flow data include spatial metadata, whereas for the bitcoin data, we used the Pearson correlation between variates to infer metadata to contextualize the prediction task. Experiments have shown that the proposed MMA technique also generalizes to these application domains (even though our motivating application is seizure prediction).

1.2 Related Work

1.2.1 Seizure Prediction and Forecasting

Epileptic seizures have four states:

1. Preictal state is a state that appears before the seizure begins marked by seizure 'aura' symptoms ranging from a few minutes up to three days.
2. Ictal state is a state that begins with the onset of seizure and ends with a seizure attack.
3. Postictal state that begins after ictal state.
4. Interictal state that starts after the postictal state of first seizure and ends before the start of preictal state of consecutive seizure.

Predicting the preictal state before the onset of seizure is very useful. Work on epileptic prediction [15] [16] [17] has been going on since a few decades using machine learning and deep learning approaches. [7] includes an extensive overview of the literature in this area. A Support Vector Machine (SVM) has been used in [18] for seizure prediction using EEG modality. In [15], authors show that the patient-specific classifier based on SVM can distinguish preictal from interictal with a high degree of sensitivity and specificity using multiple modalities like EEG and ECG. In [19], an overview of seizure detection and related prediction methods are presented using EEG and ECG and authors discuss their potential uses in closed-loop warning systems in epilepsy using SVM. Authors in [19] note the importance of studying a combination of modalities or detection technologies to interpret which yields the best results, and emphasize that these approaches may ultimately need to be individualized for patients.

In [20] authors develop a model that predicts epileptic seizures in sufficient time before the onset of seizure starts and provides a better recall. Authors have applied empirical mode decomposition (EMD) for preprocessing and have

used time and frequency domain features for training the model for the prediction task. Features extracted using a common spatial pattern (CSP) are used for training a patient-specific, linear discriminant analysis classifier in [21] using electroencephalogram (sEEG) signals. [22] looks at statistical features in high-frequency bands of interictal iEEG work in efficiently identifying the seizure onset zone in patients with focal epilepsy. Improved seizure prediction may be achieved using other external variables. There can be physiological changes observed in animals and human beings before the onset of a seizure. Changes in blood flow, blood oxygenation, and metabolism have all been shown to happen before a seizure. [23] has critically reviewed the literature on data from neocortical epilepsy using optical imaging. Optical measurements of blood flow and oxygenation may become increasingly important for predicting as well as localizing epileptic events. The combined use of electroencephalography (EEG) and functional magnetic resonance imaging (EEG-fMRI) in epilepsy is studied in [24] for posttraumatic epilepsy. In [25] authors study seizure prediction in scalp EEG using 3D convolutional neural networks with an image-based approach. Prediction of neonatal amplitude-integrated EEG based on LSTM method is studied in [26].

Multi-layer LSTM network is made use of in [27] [28] for studying the temporal context in epileptic seizure prediction. In [29] a multi-layer convolutional network is used for EEG classification tasks considering the spatial context. Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system is studied in [30] to provide interpretable decisions. Interpretable EEG seizure prediction model is given in [31] using a multi-objective evolutionary algorithm.

Seizure forecasting is a new development in terminology in [32], the strategy can be described as forecasting also since the inference of high risk periods is related to the way in which the patient functions. [33] studies the utility of digital markers, wearables, and biosensors as parameters for a seizure-forecasting algorithm. They argue that pairing up peripheral measurements to brain states could identify new relationships and insights. Another key component suggested in [33] is the diversity of the relationships in people having seizures namely seizure type and frequency indicating that pooling findings across groups is suboptimal, and that data collection will need to be done on longer time scales to allow for individualization of potential seizure-forecasting algorithms.

1.2.2 Rare Event Detection and Prediction

Rare event prediction has been studied recently in many works. In [34] authors use rare event predictive modeling for breakthrough patents. Authors have used a deep autoencoder and an anomaly detection approach to identify the most rare breakthrough patents. In [35], a Bayesian network model having causal and probabilistic semantics is used to forecast daily ozone states. Adaptive swarm balancing algorithms are studied in [36] for rare-event prediction in imbalanced healthcare data. By combining SMOTE with meta-heuristic

algorithms, authors create two methods for solving imbalanced dataset classification. In [37], authors solve the under-fitting problem for decision tree algorithms by incremental swarm optimization in rare-event healthcare classification. Rare event prediction using similarity majority under-sampling technique is studied in [38]. Deep over-sampling framework for classifying imbalanced data is introduced in [39]. Trainable undersampling technique is developed for class-imbalance learning in [40]. In [41], authors propose a new concept of rebalancing imbalanced samples in a deeply transformed latent space. Cost-sensitive learning of deep feature representations from imbalanced data is studied in [42]. Imbalanced classification via major-to-minor translation is done in [43] where less-frequent classes are augmented via translating samples (e.g., images) from more-frequent classes.

Applications in rare event prediction are studied in [44]. A real world dataset is also provided from a paper-and-pulp manufacturing industry in [44]. The dataset is a multivariate time series process. The data is extremely rare and is about a paper break event happening that commonly occurs in the paper manufacturing industry. In [45], a method is proposed to choose training data to improve the performance of deep learning models. The method represents different length multi-variate time series split into categorical variables, and measure the (dis)similarities using the distance matrix. A financial application is considered in [46] using a dynamic churn prediction framework effectively using rare event data. Variational disentanglement for rare event modeling is studied in [47].

Zero-shot learning has been studied in images extensively in the past. As part of their study in [48], authors build a Semantic Output Code (SOC) classifier for a neural decoding task and show that it can often predict words that people are thinking, from functional magnetic resonance images (fMRI) of their neural activity, without training examples for those words. Zero-shot learning through cross-modal transfer is done to recognize objects in [49]. Due to the lack of clearly expressed semantic attributes in signals, zero-shot learning is more difficult with signals. In [50], a Zero-Shot Learning (ZSL) framework is developed using signal recognition and reconstruction convolutional neural networks (SR2CNN). A combination of cross entropy loss, center loss and autoencoder loss along with a distance metric space is introduced such that semantic features have greater minimal inter-class distance than maximal intra-class distance [50].

1.2.3 Attention Mechanism In Neural Networks

A neural network is thought to be an attempt to simulate simplified human brain functions. An attempt to have deep neural networks do the same thing as humans — selectively focus on a small number of important things while disregarding others — is called Attention Mechanism. Attention mechanisms in neural networks can, at a high level, be classified as (a) self attention, (b) cross attention, and (c) externally-guided attention. *Self attention mechanisms*, such as [51], consider the patterns in the provided data itself to identify aspects

of the data to focus on. A common technique is to couple the given neural network with an encoder-decoder architecture that help identify parts of the input data that are most important to focus on [51]. Scaled dot-product self-attention is introduced by [52] in an architecture called Transformers allowing multiple attention heads for parallelization.

In [53], authors proposed that each input word should be given a certain amount of relative value in addition to being taken into consideration by a context vector; the suggested model looks for a set of points in the encoder hidden states where the most pertinent information is present everytime it generates a phrase. In [54], authors look at two kinds of attentional mechanisms, local attention that only looks at a subset of the source words at a time, and global attention that always pays attention to all source words. *Cross-attention mechanisms* are generally used in multi-variate/multi-modal machine learning tasks, where two or more separate streams of data are simultaneously analyzed [55]. These cross attention mechanisms analyze the inter-relationships between data in multiple streams to help identify what to focus on in each data stream [55]. [56] presents a framework to represent multi-scale patterns via cross-talking mechanism among multiple attention heads. Finally, *externally-guided attention* mechanisms, such as [57], take into account pre-computed saliency information, provided as a side channel, to guide which aspects of the input data to attend.

The attention model proposed in this paper LSTM-MMA leverages multi-headed attention in two layers - first to map metadata supported robust multi-variate temporal (RMT) features which are separately extracted [12] to input data and second to focus on latent semantics and LSTM output sequences. The LSTM-MMA approach can be considered as an externally-guided cross-attention mechanism, because multi-variate RMT features which are pre-extracted leveraging metadata that describe the inter-relationships among the variates is used to help guide the proposed attention mechanisms.

1.2.4 Spatio-temporal Forecasting

The proposed metadata-driven forecasting approach is related to spatio-temporal forecasting problem – in particular to those settings where the spatial relationships are encoded through a graph. Diffusion Convolutional RNN (DCRNN) [14] is a network for spatio-temporal forecasting, which relies on a graph convolution approach to take into account the spatial context or neighbors of a node (e.g. in a traffic network) using an adjacency matrix. DCRNN [14] uses the diffusion convolution operator to identify the diffusion of features for k-hops and improves the robustness of the forecasting process relying on these graph-informed features. Spatio-Temporal Graph Convolutional Networks (STGCN) [58], is another metadata-informed competitor to tackle the time series forecasting problem. STGCN [58] is a deep learning framework that leverages graph convolutional networks to capture both spatial (using an adjacency matrix) and temporal dependencies in traffic data represented as graphs. GCN-LSTM [59], a variant of STGCN is also inspired by

| | Meaning |
|---------------|--|
| \mathcal{V} | Set of variates |
| m | Number of variates |
| \mathcal{M} | Set of metadata showing variate relationships |
| T | Temporal length of multi-variate time series |
| \mathcal{Y} | Data matrix describing the multi-variate time series |
| \vec{q} | Query vector in the attention model |
| \vec{k} | Key vector in the attention model |
| \vec{v} | Value vector in the attention model |
| \vec{h} | Number of attention heads |
| \mathcal{S} | Feature scales created by the RMT algorithm |
| \mathcal{F} | RMT feature set identified in the input data |
| l | Length of the RMT feature descriptor vector |
| n_t | Number of selected RMT features covering time instance t |
| r | Target rank for feature dimensionality reduction (the reduced feature descriptor length) |
| k | Number of variates (for k -means based variate reduction) |

Table 1 Key notations

graph convolutional networks. We have adapted DCRNN [14], STGCN [58] and GCN-LSTM [59] networks as metadata-informed forecasting competitors to the proposed MMA approach.

2 LSTM-MMA : LSTM with Metadata Supported Multi-Variate Attention (MMA)

Brain seizures are rare – even in patients with post-traumatic seizure.¹ Therefore, as we discussed in the introduction, our goal in this paper is to increase the robustness of the neural nets by relying on *a priori* metadata (such as the frequency and spatial context of the sensory data streams). As described in the introduction, in this section, we propose segmented multiple LSTMs-based neural architecture with a metadata supported multi-variate attention (MMA) mechanism that leverages robust multi-variate temporal features that are fed into the neural network as a side information. In particular in this paper, we propose LSTM-MMA, which leverages available meta data such as spatial context of EEG data to extract robust, multi-variate, spatial as well as temporal features that help the neural architecture to focus on key events of the input data that are potentially relevant for rare event prediction.

1

The EEG real world dataset is imbalanced, time steps having seizure-positive labels are 7% of the total at best and < 2% at the worst (the labels are provided by expert physicians; see Section 3.1.1 for dataset details).

2.1 Meta-Data Enriched Multi-Variate Time Series Model

In this paper, we consider a metadata-enriched, multi-variate timeseries model. In particular,

a multi-variate time series is defined as a triple $\mathbf{Y} = (\mathcal{V}, \mathcal{Y}, \mathcal{M})$, where

- $\mathcal{V} = \{v_1, \dots, v_m\}$ is a set of m variates;
- \mathcal{Y} is an $T \times m$ data matrix where T is the temporal length of multi-variate time series; and
- \mathcal{M} is an application specific metadata graph that describes how the various variates in \mathcal{V} are related to each other.

Below we describe how the data matrix and metadata are constructed for the EEG data.

2.1.1 EEG Time Series and Metadata Graph

In the case of the EEG data, the multi-variate time series consists of the recorded signals from each of the sensors as depicted in [3] and is taken into consideration for analysis. Note that, depending on the system and configuration target being used, there can be 15 to 26 sensors used for different patients.

The raw EEG data that is million time stamps in length is segmented into eight second windows and power spectral density of each time window is computed by performing Fast Fourier transform on each of the individual signal segments thereby compressing the signal. The result is a multi-variate EEG time series with a total of upto 520 variates. This multi-variate time series is accompanied with a metadata graph that describe the spatial context as depicted in Figure 1 according to the International 10-20 system (C=central, T=temporal, P=parietal, F=frontal, Fp=frontal polar, O=occipital) [3]. Given these, in our experiments (reported in Section 3) for seizure onset prediction, the metadata context is captured in two alternative ways:

- *frequency context*, where we ignore the spatial context for EEG sensors, but consider the neighbor relationships among frequency channels – more specifically, the set \mathcal{V} of variates correspond to the frequency channels for each EEG sensor and the metadata graph \mathcal{M} connects neighboring channels in a given sensor to each other.
- *frequency and spatial context*, where we also consider the spatial placement of the EEG sensors – in this case, we have the same set of variates, but not only the neighboring frequency channel variates corresponding to the same sensor are connected to each other in the metadata graph \mathcal{M} , but also frequency channel variates corresponding to the same sensor in neighboring sensors according to the spatial context depicted in Figure 1 are connected to each other.

These two contexts *frequency context* and *frequency and spatial context* are used for direct learning and transfer learning approaches for LSTM-MMA as explained in 3.1.1.

2.2 Robust Multi-Variate Temporal Features

Our key argument in this paper is that multi-variate time series carry robust localized multi-variate temporal and spatial features that could help predict critical events; however, the lack of sufficient data to train for these events makes it impossible for neural architectures to identify and make use of these features. We therefore, propose that these features are identified through a process external to the neural network architecture [12] and then used as a side information to train the neural network.

2.2.1 Overview of the Metadata Supported RMT Extraction Process

In this paper, we rely on the metadata supported robust multi-variate temporal (RMT) feature extraction algorithm proposed in [12]. Intuitively, a RMT feature is a fragment of a multi-variate time series that is maximally different from its immediate neighborhood, both in time and across variate relationships specified by the metadata as shown in [12]. Multi-variate temporal features of interest can be of different lengths and may cover different number of variates.

As shown in [12], the Gaussian smoothing process is guided by a metadata graph, which captures the variate relationships (e.g. defined by the spatial context for EEG data) – and a scale space is constructed through iterative smoothing of both the time series and the metadata graph in order to locate such features of different sizes. This creates different resolution versions of the input data and, thus, helps identify features with different amount of details in time and in spatial context (in terms of the number of variates involved). We denote the set of scales, each corresponding to a different temporal; feature size, created by this process with \mathcal{S} .

Next, the process identifies candidate features of interest across multiple scales of the given multi-variate time series by searching over multiple scales and variates of the given series. Each candidate RMT feature that is not poorly localized has a *temporal-scope* (a beginning and an end in time) and a *variate-scope* (a set of variates involved in the feature). These candidate features of interest are those with the largest variations with respect to their neighbors in time, variates, and scale.

At the following step, those candidate features that are poorly localized (and hence are inappropriate to use as key events) are eliminated.

The above process leads to a set, \mathcal{F} , of RMT features where each feature, $f_i \in \mathcal{F}$, extracted from \mathbf{Y} , is a pair of the form, $f_i = \langle pos_i, \vec{d}_i \rangle$:

Here, $pos_i = \langle v_i, t_i, s_i \rangle$ is a VTS triple denoting the position of the feature in the scale-space of the multi-variate time series, where v_i is the index of the variate at which the feature is *centered*, t_i is the time instant around which

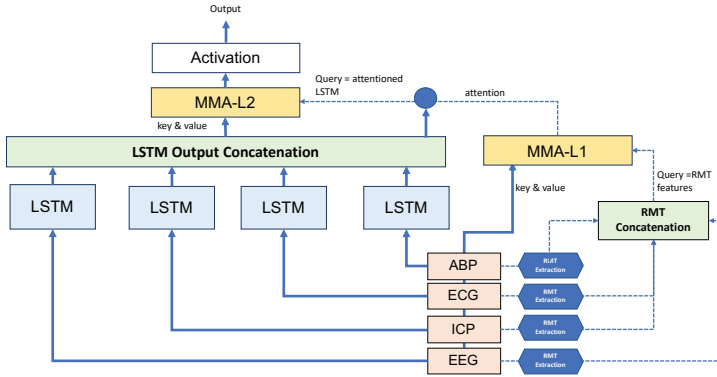


Fig. 2 (Multi-modal) LSTM-MMA model which enhances traditional LSTM neural network architecture with a metadata supported, multi-variate attention (MMA) mechanism

the duration of the feature is *centered*, and $s_i \in \mathcal{S}$ is the temporal/variant smoothing scale in which the feature is identified; and note that this triple also defines the *temporal and variates scopes* of the RMT feature.

\vec{d}_i is a descriptor vector, representing a gradient histogram describing the temporal structure (in terms of the distribution of local gradients) corresponding to the identified key event. Note that the above approach to identify RMT features has several advantages as mentioned in [12].

In addition to being robust against noise and transformations such as temporal shifts, dropped/missing variates, the identified salient features have scale invariance which enables multi-resolution analysis. The temporal and spatial scales at which a multi-variate feature is located give an indication about the scope both in terms of duration and the number of variates involved of the multi-variate feature. The value of s_i is the temporal/spatial scope of the key event corresponding to the RMT feature. In particular, since we use Gaussian smoothing to obtain the scale-space, each scale s_i has a corresponding Gaussian smoothing parameter, σ_i , and the temporal scope of the feature is $6\sigma_i$ since $3\sigma_i$ from the center point t_i , in both directions, would cover approximately 99.73% of the contributions to the smoothing.

2.2.2 EEG Data and RMT Features

In the case of EEG data, we consider the connectivity graph, \mathcal{M} , outlined in Section 2.1.1, that considers neighborhoods in frequency and spatial contexts.

The input data (after Fast Fourier transformation) from all EEG sensors at a particular time segment are concatenated and RMT features are extracted for that time segment. These are then fed into the proposed LSTM-MMA model as described next.

2.3 Metadata Supported Multivariate Attention (MMA)

In this section, we develop a LSTM-MMA model which enhances traditional LSTM neural network architecture with a metadata supported, multi-variate attention (MMA) mechanism (Figure 2).

In particular, a multi headed attention unit has been used in the model inspired by the transformer from [52] to operate on input data \mathcal{Y} , along with the RMT features extracted from this \mathcal{Y} . Intuitively, the multi headed attention maps a query and set of key-value pairs to an output. The query vector \vec{q} represents the inference question, the key \vec{k} represents the available context information, and a value vector \vec{v} specified the values on which the attention is applied. The attention matrix is constructed through the dot product of all keys and queries, normalized via softmax, to create a mapping of elements in the key sequence corresponding to the data needed for each query. After taking softmax, the normalized attention matrix is applied on the value vector. More specifically, given query, key, and value vectors, \vec{q} , \vec{k} , and \vec{v} , respectively, we have

$$\text{Attention}(\vec{q}, \vec{k}, \vec{v}) = \text{softmax}\left(\frac{\vec{q}\vec{k}^T}{\sqrt{d_k}}\right)\vec{v}, \quad (1)$$

where d_k is the length of the key vector \vec{k} .

2.3.1 First Metadata Supported Multi-Variate Attention Layer

Intuitively, the first attention layer of LSTM-MMA helps the multi-modal LSTM model to focus on different parts of the combined EEG, ICP, ECG and ABP data, as a function of the concatenated RMT features corresponding to each time step. Therefore, in the first layer, the query vector, \vec{q} is the RMT descriptors extracted from the input data.

The key, \vec{k} , and value, \vec{v} , vectors both are set to be the input multi-variate time series.

With respect to a given time instance, t can be within the scopes of multiple RMT features. As shown in Figure 3, the time instance t is covered by multiple RMT features. Nevertheless, the distance of t to the centers of these features may be different, therefore its contribution to these features may vary. To account for this, for each feature f_* that covers t in its scope, we compute a contribution value

$$\text{contrib}(t, f_*) = e^{-\frac{1}{2}\left(\frac{t_*-t}{\sigma_*}\right)^2}, \quad (2)$$

which captures the Gaussian nature of the smoothing process applied to obtain the features. Note that, since the $\text{contrib}(t, f_*)$ takes a value between 0 and 1, it can be treated also as a probability of contribution. Therefore, to identify a set of features, \mathcal{F}_t , that correspond to time instance t , we randomly select n_t RMT features based on the individual contribution probabilities of the features covering t . Let us denote the length of the RMT feature descriptor vector with l (in our experiments $l = 128$). In LSTM-MMA, for each time instance t , we stack the n_t many RMT feature descriptors corresponding to features in \mathcal{F}_t ,

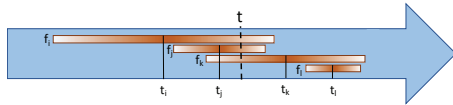


Fig. 3 Three features, f_i , f_j , f_k , and f_l centered around time instances, t_i , t_j , t_k , and t_l respectively – note that the scopes of the RMT features are defined by the Gaussian smoothing parameters (σ_i , σ_j , σ_k and σ_l corresponding to each feature); the time instance t is within the scopes of the first three of these four features, but since t is closest to t_j its contribution is highest relative to the feature f_j

constructing a data structure (a matrix, M_t) of size $n_t \times l$. The above process is done individually for each of the data sources/modalities to get matrix M_t which is then fed into LSTM-MMA to support attention at time t .

2.3.2 Second Metadata Supported Multi-Variate Attention Layer

The first attention layer helps LSTM-MMA to focus on different latent semantics, as a function of the RMT features. Therefore in the second layer, the query vector, \vec{q} , is the output of the LSTM models combined with the attention weights from first layer; whereas the key, \vec{k} , and value vectors, \vec{v} , are the LSTM output sequences, each with its own descriptive vector.

For the second attention layer, we are using a MultiHead attention unit as it allows the model to jointly attend to information from the different latent subspaces. Each attention head is of the form

$$head_i = Attention(\vec{q}W_{Q,i}, \vec{k}W_{K,i}, \vec{v}W_{V,i}), \quad (3)$$

where $W_{Q,i}$, $W_{K,i}$ and $W_{V,i}$ are the weights corresponding to the query, key and value vectors, respectively.

MultiHead Attention from the transformer model [52] is applied as follows for our purposes in both attention layers. MultiHead Attention is subdivided into Pre and Post Attention modules so we have a handle to the Attention matrix itself for explainability purposes. Pre-Attention module returns the normalized Attention matrix after softmax operation on the dot product of all keys and queries. Value vector is then taken through a linear transformation via a fully connected layer. Then in the Post-Attention module a dot product takes place between the Attention matrix and linearly transformed Value vector.

Given the query, key and value vectors, an h -headed model is trained by considering

$$MultiHead(\vec{q}, \vec{k}, \vec{v}) = [head_1; \dots; head_h]W_O, \quad (4)$$

where W_O captures the weights for the overall output. In our experiments, the number of heads is set to eight as in the paper [52] in the first and second layer for the EEG multi-modal dataset.

2.3.3 L1 Regularized MultiHead Attention

We also trained a variant of the transformer model with weights to the attention heads namely Least Absolute Shrinkage and Selection Operator (Lasso) regularization (L1) [60]. In this variant we have an additional parameter in the Post-Attention module for learning the weights on a head. Relevant heads in the MultiHead Attention are paid more attention and less relevant heads are not selected using Lasso (L1) regularization. The learned weights of the heads are applied to the output of the dot product between the Attention matrix and Value vector.

$$MultiHead(\vec{q}, \vec{k}, \vec{v}) = [W_{H,1}head_1; \dots; W_{H,h}head_h]W_O, \quad (5)$$

Note that, as we see in Figure 2 at the final step, the output of the dual attention layer goes through a final activation step to complete the inference process: *sigmoid* activation is used for binary (“no-event”, “event”) classification, whereas for regression tasks, we have used mean squared error metric.

2.4 Noise Reduction in Multi-modal RMT Features used for Attention

The process for RMT feature extraction described in Section 2.2 leads to a descriptor vector for each RMT feature – the descriptor size² must be selected in a way that reflects the temporal characteristics of the time series; if a multi-variate time series contains many similar features, it might be more advantageous to use large descriptors that can better discriminate: these large descriptors would not only include information that describe the corresponding features, but would also describe the temporal contexts in which these features are located.

2.4.1 PCA-based Reduction of RMT feature descriptors

As described in the previous section, in LSTM-MMA, we stack multiple RMT feature descriptors corresponding to each time step for a modality leading to a data structure (matrix) M_t for each time instant t . While this structure can be fed as is to the attention mechanism, we note that due to its size and noise inherent in the feature extraction process, this may not be a very effective strategy. We instead consider noise elimination and dimensionality reduction of the RMT feature descriptors before they are fed into the attention process. This is done for each data channel of a modality once using Principal Component Analysis (PCA) with a user provided target rank r , before the stacking operation.

In PCA based approach, the input is an $a \times b$ matrix M_t , where $a = n_t$ is the number of RMT feature descriptors in a data channel of a particular modality and $b = l$ is the length of the RMT feature descriptor vector. We first

²In experiments reported in Section 3, the descriptor vector length is 128.

obtain the corresponding $a \times a$ covariance matrix C_t , which is then decomposed into $C_t = U_t \Sigma_t U_t^T$, where the $a \times c$ matrix U_t records the c eigenvectors and diagonal matrix Σ_t records the corresponding eigenvalues of the matrix C_t . Given a target rank $r \leq c$, we then decompose C_t as $\hat{C}_t = U_t' \Sigma_t' U_t'^T$ where the $a \times r$ matrix U_t' records the r eigenvectors and diagonal matrix Σ_t' records the corresponding eigenvalues of the matrix \hat{C}_t with target rank r . The matrix U_t' is used as input instead of matrix M_t for the stacking operation.

Finally the reduced and stacked RMT feature descriptors of each modality are concatenated, that is RMT feature descriptor data from each channel/sensor of EEG along with ICP, ECG and ABP modalities are concatenated and then fed into the attention mechanism as the query matrix. Note in case of frequency and spatial context, RMT feature data from all EEG time segments are concatenated as a single channel before reducing and stacking operations.

2.5 Variate Reduction using Adaptive Clustering

We also consider an additional variate reduction strategy to complement the learning process. In particular, we apply k -means clustering to the input data to reduce the number of variates from m to k . The clustering is applied on the variates in the combined time series data of all the EEG data channels, ICP, ECG and ABP channels. After the clusters are obtained under the Euclidean distance model, the resulting k cluster centroids are used to construct the data matrix passed to the first layer of LSTM-MMA (note that the RMT features used for attention are extracted directly from the original data matrix before the variate reduction). Optimum k clusters were learned using Scikit-learn's [13] silhouette score method, which is an indicator of the quality of a cluster. Note that the EEG data might have difference in density with ICP, ECG and ABP data, that is EEG may be packed more loosely than others for some patients. In such cases a trial and error method is utilized to choose the next larger k given by the silhouette score method [13].

3 Experiments

In this section, we present experiment results to evaluate the effectiveness of LSTM-MMA (LSTMs with metadata supported multi-variate attention) in predicting onsets in multi-variate multi-modal time series. (a) Since our motivating application is seizure prediction, the primary data set we use is EEG data, complemented with other physiological data sources, including ICP, ECG and ABP, with rare seizure events labeled by physicians. Since the data set cannot be released due to HIPAA protections, in order to illustrate the broader applicability, reproducibility, and generalizability of the proposed techniques, we also evaluate LSTM-MMA in other rare event prediction and forecasting tasks, namely (b) anomaly prediction in COVID data for the different states in United States, (c) traffic flow forecasting, and (d) bitcoin price forecasting.³

³Since the healthcare data is HIPAA protected, we make the code available. Also we have publicly available multi-modal COVID, traffic, Bitcoin and S&P index datasets and code for reproduction of results at <https://rb.gy/umbzt8>.

| LSTM-MMA and baseline Hyperparameters(using Keras 2.3.1) | Value |
|--|-----------|
| Batch size for seizure | 60 |
| Epochs for seizure classification task | ≤ 17 |
| Hidden nodes of LSTM for EEG | 100 |
| Hidden nodes of LSTM for ICP, ECG and ABP | 20 |
| Batch size for COVID | 60 |
| Epochs for COVID regression task | 150 |
| Hidden nodes of LSTM for COVID | 100 |
| L1 regularization penalty for COVID | 0.001 |
| Batch size for Traffic | 10 |
| Epochs for Traffic regression task | ≤ 20 |
| Hidden nodes of LSTM for Traffic | 100 |
| L1 regularization penalty for Traffic | 0.0001 |
| Batch size for Bitcoin | 16 |
| Epochs for Bitcoin regression task | 10 |
| Hidden nodes of LSTM for Bitcoin | 8 |
| L1 regularization penalty for Bitcoin | 0.00001 |
| Learning rate(Adam optimizer) for all datasets | 0.001 |
| Number of LSTM-MMA attention heads for EEG, Bitcoin, Traffic (h) | (8,8) |
| Number of LSTM-MMA attention heads for COVID (h) | (8,52) |

| RMT Hyperparameters | Value |
|--|-----------------------|
| Smallest scope | ~ 60 time units |
| Largest scope | ~ 420 time units |
| Scales for freq. context for EEG ($-S''$) | 12 |
| Scales for freq. & spat. context for EEG($-S''$) | 3 |
| Scales for ICP, ECG and ABP ($-S''$) | 3 |
| Descriptor length (l) | 128 |
| Reduced descriptor length with PCA (r) | 10 |

Table 2 Default hyperparameters

As mentioned in Section 1.1, two additional datasets namely COVID and traffic datasets evaluate spatial context as metadata and bitcoin dataset has correlation of variates using Pearson correlation [61] as metadata.

Unless specified otherwise, the experiments are conducted using the default hyperparameter values in Table 2. MacBook Pro with Intel UHD Graphics 630 1536 MB and Linux machines (Ubuntu 18.0)⁴with GPU 16GB RAM were used for experiments.

3.1 Evaluation Scenarios

Here we describe the four multi-variate data sets we have considered in these experiments, along with the supporting metadata and the prediction and forecasting tasks. Evaluation usecases are mainly LSTM-MMA with spatial context (explicit) and without spatial context (implicit). Adaptive Clustering evaluation usecases are done for EEG, ICP, ECG and ABP datasets for variate reduction as mentioned in Section 2.5. For COVID, traffic and bitcoin datasets which have smaller number of variates when compared to EEG multimodal

⁴Provided by NSF testbed “Chameleon: A Large-Scale Re-configurable Experimental Environment for Cloud Research”.

dataset, we evaluate spatial context with and without L1 (Lasso) regression (instead of clustering) which selects those features that are useful. We also provide alternative baselines/competitors against LSTM-MMA, as explained in Section 3.2 for comparison of results.

3.1.1 Seizure Data and Seizure Onset Prediction Task

We first describe the seizure data sets, EEG, ICP, ECG and ABP. The data set is partitioned into a training set, validation set, and test set, with 60%, 20%, and 20% of the original data each, respectively. In order to ensure that each region namely training, valid and test has similar distribution of positive and negative labels, the time series are chunked and these chunks are shuffled in a way that preserves the rate of positive labels in each of the three regions.

Seizure Onset

In these experiments, the seizure onset is defined as the first 48 seconds (6×8 time units) of the seizure event.

Seizure Onset Prediction Task

The seizure onset prediction task is defined as identifying a seizure onset occurrence between 4.4 (33 time units) to 5.1 (38 time units) minutes ahead of the time.

EEG Seizure Dataset

The first set of experiments were performed on the EEG dataset provided by Phoenix Children's. The dataset records EEG time series and seizure events, marked by physicians, for 19 patients with 86 seizure events. There are two types of patients - patients with patterns of seizure and patients with single or multiple clusters of seizure events. These patients have approximately 7% anomalies at most. There are also several patients marked with very low percent, 2% or less anomalies with 0.5% anomalies at worst. Five specific patients each from a particular category namely "ping-pong" seizure (seven seizure events), "single long seizure" cluster having one single event of seizure, "multiple seizure" clusters - one patient with 7% anomalies (three seizure events) and another one with 2% or less anomalies (four seizure events), lateral seizures which have both lateral patterns and seizure clusters (seven seizure events) are chosen for reporting purposes. There are a total of 22 seizure events reported for these 5 patients. The data set contains EEG recordings of 8 second windows upto 106,000 windows. The raw EEG data were recorded from 26 channels with a sampling rate of 256 Hz, using both referential and bipolar montage. While the sensor readings are used directly in referential montage, in bipolar montage the signals are differenced according to a spatial connectivity graph and the differenced data are used instead of the original readings. The EEG time series are segmented into eight second windows and, for each window, the corresponding power spectral density, with 20 frequency bands, is computed

using Fast Fourier transform (0 to 19 Hz with 1 Hz bins). This leads to a time series with $(26 \times 20) = 520$ variates and $(216 \times 10^6 \div (256 \times 8)) = 105944$ time steps.

In Section 2.1.1, we described the (a) frequency and (b) frequency and spatial metadata used for implementing metadata supported multi-variate attention (MMA) on EEG data. The time series were chunked into sequences of length 500 for training the LSTM (default unless specified). For patients with very short seizures (2% and lower anomalies), the chunk lengths were reduced to 50 instead of using 500 as they had zero anomalous samples in the training region to do a threeway test-train split with sequence length of 500. For the frequency and spatial context scenario, time segment length of 500 is used for RMT feature extraction (default unless specified). The chunks were shuffled in such a way that training, validation, and testing sets have similar ratios of events.

ICP, ECG, and ABP Datasets

The Intra Cranial Pressure (ICP) data were recorded with a sampling rate of 125 Hz. ICP data also have 8 second windows like EEG data and power spectral density, with 20 frequency bands (0 to 7.6 Hz with 0.4 Hz bins) is computed using Fast Fourier Transform. The electrocardiogram (ECG) data were recorded with a sampling rate of 500 Hz. ECG data also have 8 second windows like EEG and ICP data and power spectral density, with 20 frequency bands (0 to 7.6 Hz with 0.4 Hz bins) are computed using Fast Fourier Transform. The Artrial blood pressure (ABP) data were recorded with a sampling rate of 125 Hz. ABP data also have 8 second windows like ICP and ECG data and power spectral density, with 20 frequency bands (0 to 7.6 Hz with 0.4 Hz bins) are computed using Fast Fourier Transform.

Metadata

In Section 2.1.1, we described the frequency and spatial contexts used for implementing metadata supported multi-variate attention (MMA) on EEG sensor data.

Unlike EEG data, ICP, ECG and ABP data sets do not have spatial context.

Model Transfer Across Patients

In this section, we consider both direct and transfer learning scenarios. In direct learning the same patient's data (all 19 patients) are used for both training and testing. In transfer learning scenarios, model trained with one patient's data is used for predicting onsets for one another patient. The patient with "ping-pong" seizure having seven seizure events is chosen as donor as it shows highest pattern diversity. Rest of the 19 patients are test patients (having a total of $86 - 7 = 79$ seizure events) in transfer learning scenarios. For reporting purposes, we have chosen five patients from particular categories as mentioned earlier - "ping-pong" seizure patient who is the donor and four test patients

for transfer learning (reporting a total of $22 - 7 = 15$ seizure events for transfer learning) - as the proposed methods work similarly for patients belonging in a particular category. Note that when the test patient is a patient with very short seizures (2% and lower anomalies), the donor/provider is also trained with chunk lengths of 50 instead of 500.

The input data from the EEG sensors of the donor/provider patient are given as input along with ICP, ECG and ABP to multiple (four) segmented/separate LSTMs. In data preprocessing, different versions or combinations of multi-modal data (both for freq. context and freq. and spat. context) are made by masking to enable learning, so as to learn which of the modalities are strong predictors. The model learns to make predictions on both full and partial data simultaneously to get results. Based on the Critical Onset F1 Score described in Accuracy Measures 3.1.1, ranking can be done between the four sources of variates from EEG, ICP, ECG and ABP modalities.

Accuracy Measures

For the seizure onset prediction tasks, we assess the accuracy of different models using the F1-score metric (i.e., harmonic means of recall and precision):

$$F1Score = \frac{(2 \times Recall \times Precision)}{(Recall + Precision)} \quad (6)$$

Here, *Recall* is the ratio between the number of positive samples correctly classified as positive (True Positive) by the model to the sum of true positives and true samples falsely predicted as negatives (True Positive + False Negative). *Precision* is the ratio between the number of positive samples correctly classified as positive (True Positive) by the model to sum of the true positives and negative samples falsely predicted as positives (True Positive + False Positive). For the EEG, ICP, ECG and ABP datasets, we have defined two additional accuracy metrics for early prediction task which is a binary classification task. Critical Onset Recall and Critical Onset Precision accuracy metrics are calculated in the critical region between non-seizures and onset region for early prediction of seizure onset. The Onset region is our positive class and Non-Seizure region is our negative class for seizure early prediction task.

The Critical Onset Recall, Critical Onset Precision and Critical Onset F1 Score are defined as follows:

$$CriticalOnsetRecall = \frac{TrueOnset}{(TrueOnset + FalseNonSeizure)} \quad (7)$$

$$CriticalOnsetPrecision = \frac{TrueOnset}{(TrueOnset + FalseOnset)} \quad (8)$$

Each experiment has been executed a minimum of 10 times and we compute modified Recall and Precision namely Critical Onset Recall and Precision as per equations 7 and 8 for each run. Critical Onset Recall and Precision values are then used to calculate Critical Onset F1 score using equation 6 for

each run. The mean of these accuracy measures namely Critical Onset Recall, Precision and micro F1 scores from the 10 runs is computed for reporting purposes. For each run we use an adaptive threshold cutoff(s) strategy where the Critical Onset F1 score is best for the imbalanced dataset. This is because the traditional threshold of 0.5 is not found to be suitable for imbalanced datasets, the threshold is observed to be way lower.

3.1.2 COVID Dataset and Prediction of Rate of Change

The second task has two goals, predicting the rate of change of log of cases and rate of change of log of deaths for 52 US states for the next day.

For these experiments, COVID data from January 21st 2020 till the peak month of March 2021 (upto February 28th 2021) in [62] joined with an external data namely demographic data [63] and was processed to get 5 variates namely (a) log of cases, (b) log of deaths, (c) days from the beginning of the dataset, (d) rate of change of log of cases, and (e) rate of change of log of deaths. Sequence/chunk length used for segmenting this dataset is 20. The time segment length for RMT spatial feature extraction is also 20. The data set is partitioned into a training set, validation set, and test set, with 60%, 20%, and 20% of the joined data.

Metadata

The metadata for the COVID prediction task is a graph, where each node is one of the 5 variates for one of the 52 US states and two nodes corresponding to the same variate have an edge between them if the corresponding states are neighbors.

Accuracy Measures

For the COVID prediction task (which requires a regression model, rather than a classification model), we report root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE). Due to very low MSE values, four positions after the decimal is used here.

3.1.3 Traffic Flow Forecasting

Our third data set focuses on a regression task for traffic prediction based on modeling spatial and temporal dynamics in road networks, to predict the travel time for a given departure time. The dataset is provided by Highways England [64]. This dataset offers the average travel/journey time for 15-min time intervals starting in April 2009 on all motorways and "A" roads in England that are under the control of the Highways Agency, often known as the Strategic Road Network. On motorways and 'A' routes, information about speed and traffic flow is also provided, along with the average journey time at 15-min intervals. There are 96 distinct possible departure times because there are 96 time periods in a day. The dataset's journey times were derived using GPS-based real-world vehicle observations. In this section, we consider

31 day period for the month of January 2011, the length of the time series being $31 \times 96 = 2976$ time steps. The time series were split into sequences of length 4 for training. The multi-modal inputs to the LSTM model are the previous travel times and departure times from multiple sensors on same road and different roads. There is an embedding layer used for representing time for regression problem, this is to learn traffic congestion similarities between previous timestamps and the query timestamp. In the experiments we look at the previous 4 travel times (one hour history) to predict the next travel time (15 minutes). We have selected the highways/roads A11 with Hatris link sensors 'AL2272', 'AL2270', 'AL2844' as neighboring sensors and another road A1 with sensor 'AL1165A' for the travel time study. There are 20 ($5 * 4$) variates in traffic dataset: travel time, day type, total traffic flow, average speed, and quality index. We are forecasting the traffic flow time for 'AL2272' based on the neighborhood graph.

Metadata

The metadata for the traffic flow forecasting task is a graph, where each node is one of the 5 variates for one of the sensors in a road in England and two nodes corresponding to variates of two sensors in the same road have an edge between them as the corresponding sensors are neighbors and 0 if they are not on the same road according to [64]. The training set is 70%, valid set 20%, test set 10% of the input data to include outliers in training data.

Accuracy Measures

For the traffic flow forecasting task (which requires a regression model, rather than a classification model), we report root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE).

3.1.4 Bitcoin Price Forecasting

Bitcoin is one type of cryptocurrency that has been steadily increasing over the past several years, with occasional abrupt drops that have no apparent impact on the stock market. Due to the constant fluctuations it would be good to learn to forecast the bitcoin price. As training datasets, we used the historical price of bitcoin (BITCOIN_USD or BTC_USD) and S&P 500 index of US companies from Yahoo website ⁵. The respective costs for both of them are listed on the Yahoo financial website and are expressed in US dollars. It has a Date timestamp, the value at Open, High, Low, Closing price, Adjusted Closing price and the volume traded in Bitcoin and USD. We use the normalized value at Open, High, Low, Closing price, Adjusted Closing price and the volume as the predictors. We are analyzing the price of bitcoin for the time period from January 2nd 2018, to July 29th 2022. We are forecasting using a sliding window starting from $p = 30th$ day onwards.

⁵<https://finance.yahoo.com/quote/BTC-USD?p=BTC-USD>

Metadata

The metadata for the bitcoin price forecasting task is a graph, where each node is one of the 6 variates for one of the two assets bitcoin and S & P 500. All corresponding 6+6 variates of the two assets are given a score in the metadata matrix using Pearson correlation coefficient which measures the linear correlation between two variates [61].

Accuracy Measures

For the bitcoin price forecasting task (which requires a regression model, rather than a classification model), we report root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE). Due to very low MSE values, four positions after the decimal is used here.

| Patient with "Ping-Pong" Seizures | | | |
|--|------------------|---------------------|------------------|
| Model | Mean CORecall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 0.83 | 0.91 | 0.87 |
| LSTM w/o var. cluster; no context | 0.63 | 0.23 | 0.34 |
| LSTM w/o var. cluster; no context; with self-attention | 0.68 | 0.26 | 0.38 |
| CNN w/o var. cluster; no context | 0.42 | 0.13 | 0.20 |
| CNN w/o var. cluster; no context; with self-attention | 0.83 | 0.58 | 0.68 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.53 | 0.03 | 0.06 |
| STGCN w/o var. cluster; freq. & spat. context | 0.11 | 0.02 | 0.03 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.83 | 0.08 | 0.14 |
| ICP | | | |
| LSTM-MMA w/o var. cluster; freq. context (spat. context N/A) | 1.00 | 0.50 | 0.67 |
| LSTM w/o var. cluster; no context (spat. context N/A) | 0.23 | 0.06 | 0.10 |
| LSTM w/o var. cluster; no context; with self-attention (spat. context N/A) | 0.70 | 0.12 | 0.20 |
| CNN w/o var. cluster; no context (spat. context N/A) | 0.48 | 0.04 | 0.07 |
| CNN w/o var. cluster; no context; with self-attention (spat. context N/A) | 0.60 | 0.15 | 0.24 |
| DCRNN w/o var. cluster; freq. context (spat. context N/A) | 0.90 | 0.25 | 0.39 |
| STGCN w/o var. cluster; freq. context (spat. context N/A) | 0.06 | 0.06 | 0.06 |
| GCN-LSTM w/o var. cluster; freq. context (spat. context N/A) | 0.22 | 0.03 | 0.05 |
| EEG and ICP | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 0.87 | 1.00 | 0.93 |
| LSTM w/o var. cluster; no context | 0.52 | 0.29 | 0.37 |
| LSTM w/o var. cluster; no context; with self-attention | 0.63 | 0.37 | 0.47 |
| CNN w/o var. cluster; no context | 0.39 | 0.08 | 0.13 |
| CNN w/o var. cluster; no context; with self-attention | 1.00 | 0.13 | 0.23 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.70 | 0.09 | 0.16 |
| STGCN w/o var. cluster; freq. & spat. context | 0.06 | 0.06 | 0.06 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.83 | 0.06 | 0.11 |
| EEG, ICP and ECG | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| LSTM w/o var. cluster; no context | 0.42 | 0.45 | 0.43 |
| LSTM w/o var. cluster; no context; with self-attention | 0.85 | 0.35 | 0.50 |
| CNN w/o var. cluster; no context | 0.43 | 0.02 | 0.04 |
| CNN w/o var. cluster; no context; with self-attention | 1.00 | 0.14 | 0.25 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.75 | 0.04 | 0.08 |
| STGCN w/o var. cluster; freq. & spat. context | 0.06 | 0.11 | 0.08 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.22 | 0.20 | 0.21 |
| EEG, ICP, ECG and ABP | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| LSTM w/o var. cluster; no context | 0.47 | 0.34 | 0.39 |
| LSTM w/o var. cluster; no context; with self-attention | 0.75 | 0.34 | 0.47 |
| CNN w/o var. cluster; no context | 0.46 | 0.02 | 0.04 |
| CNN w/o var. cluster; no context; with self-attention | 0.58 | 0.09 | 0.16 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.90 | 0.01 | 0.02 |
| STGCN w/o var. cluster; freq. & spat. context | 0.11 | 0.11 | 0.11 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.67 | 0.04 | 0.08 |

Table 3 Comparison of LSTM-MMA against baselines for direct learning freq. & spat. context using EEG, ICP, ECG and ABP data (5.7 % rare event) – PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

| Patient with Single Long Seizure | | | |
|--|-----------------|---------------------|------------------|
| Model | Mean CORcall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 0.92 | 0.06 | 0.11 |
| LSTM w/o var. cluster; no context | 0.53 | 0.02 | 0.04 |
| LSTM w/o var. cluster; no context; with self-attention | 0.60 | 0.06 | 0.11 |
| CNN w/o var. cluster; no context | 0.88 | 0.01 | 0.02 |
| CNN w/o var. cluster; no context; with self-attention | 0.80 | 0.05 | 0.09 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.63 | 0.02 | 0.04 |
| STGCN w/o var. cluster; freq. & spat. context | 0.60 | 0.01 | 0.02 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.83 | 0.02 | 0.04 |
| ICP | | | |
| LSTM-MMA w/o var. cluster; freq. context (spat. context N/A) | 1.00 | 1.00 | 1.00 |
| LSTM w/o var. cluster; no context (spat. context N/A) | 0.53 | 0.78 | 0.63 |
| LSTM w/o var. cluster; no context; with self-attention (spat. context N/A) | 1.00 | 1.00 | 1.00 |
| CNN w/o var. cluster; no context (spat. context N/A) | 0.90 | 0.99 | 0.94 |
| CNN w/o var. cluster; no context; with self-attention (spat. context N/A) | 0.97 | 0.97 | 0.97 |
| DCRNN w/o var. cluster; freq. context (spat. context N/A) | 1.00 | 1.00 | 1.00 |
| STGCN w/o var. cluster; freq. context (spat. context N/A) | 0.83 | 0.83 | 0.83 |
| GCN-LSTM w/o var. cluster; freq. context (spat. context N/A) | 0.83 | 1.00 | 0.91 |
| EEG and ICP | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 0.99 | 0.87 | 0.93 |
| LSTM w/o var. cluster; no context | 0.39 | 0.59 | 0.47 |
| LSTM w/o var. cluster; no context; with self-attention | 0.80 | 0.35 | 0.49 |
| CNN w/o var. cluster; no context | 0.50 | 0.10 | 0.17 |
| CNN w/o var. cluster; no context; with self-attention | 0.95 | 0.76 | 0.84 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.77 | 0.20 | 0.32 |
| STGCN w/o var. cluster; freq. & spat. context | 0.83 | 0.50 | 0.62 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.17 | 0.50 | 0.25 |
| EEG, ICP and ECG | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 1.00 | 0.90 | 0.95 |
| LSTM w/o var. cluster; no context | 0.38 | 0.35 | 0.36 |
| LSTM w/o var. cluster; no context; with self-attention | 0.90 | 0.39 | 0.54 |
| CNN w/o var. cluster; no context | 0.31 | 0.13 | 0.18 |
| CNN w/o var. cluster; no context; with self-attention | 1.00 | 0.58 | 0.73 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.68 | 0.24 | 0.35 |
| STGCN w/o var. cluster; freq. & spat. context | 0.50 | 0.13 | 0.21 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.33 | 0.67 | 0.44 |
| EEG, ICP, ECG and ABP | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 1.00 | 0.92 | 0.96 |
| LSTM w/o var. cluster; no context | 0.27 | 0.38 | 0.32 |
| LSTM w/o var. cluster; no context; with self-attention | 0.90 | 0.47 | 0.62 |
| CNN w/o var. cluster; no context | 0.28 | 0.17 | 0.21 |
| CNN w/o var. cluster; no context; with self-attention | 0.80 | 0.36 | 0.50 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.39 | 0.13 | 0.20 |
| STGCN w/o var. cluster; freq. & spat. context | 0.50 | 0.14 | 0.22 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.33 | 0.67 | 0.44 |

Table 4 Comparison of LSTM-MMA against baselines for direct learning freq. & spat. context using EEG, ICP, ECG and ABP data (7 % rare event) – PCA reduction applied by default on RMT attention (the higher, the better)– $p_{max} = 5.1m$, $p_{min} = 4.4m$.

3.2 Alternative Baselines

We have chosen LSTM [28] [65], CNN [29], Diffusion Convolution RNN (DCRNN [14]), Spatio Temporal Graph Convolution Network (STGCN [58]) and GCN-LSTM [59] as state of the art baselines for considering temporal and spatial aspects for onset prediction and forecasting tasks. More specifically, we consider 7 competitors against the proposed MMA based attention mechanism: **Competitor #1 (Vanilla LSTM)**. Four segmented/separated (one for EEG, ICP, ECG and ABP) LSTM layers without context or MMA (Metadata Supported Multi-variate Attention) is the first baseline we are using. All LSTM based models have the intermediate outcomes (in Keras this is the `return_sequence=True` parameter turned on which makes it possible to access the hidden state output for each input time step) processed as opposed to looking only at the last output state.

| Patient with Multiple Seizures | | | |
|--|-----------------|---------------------|------------------|
| Model | Mean CORcall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 0.92 | 0.23 | 0.37 |
| LSTM w/o var. cluster; no context | 0.82 | 0.08 | 0.15 |
| LSTM w/o var. cluster; no context; with self-attention | 0.40 | 0.13 | 0.20 |
| CNN w/o var. cluster; no context | 0.72 | 0.19 | 0.30 |
| CNN w/o var. cluster; no context; with self-attention | 0.60 | 0.22 | 0.32 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.92 | 0.13 | 0.23 |
| STGCN w/o var. cluster; freq. & spat. context | 0.33 | 0.06 | 0.10 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.17 | 0.14 | 0.15 |
| ICP | | | |
| LSTM-MMA w/o var. cluster; freq. context (spat. context N/A) | 1.00 | 0.50 | 0.67 |
| LSTM w/o var. cluster; no context (spat. context N/A) | 0.97 | 0.22 | 0.36 |
| LSTM w/o var. cluster; no context; with self-attention (spat. context N/A) | 0.90 | 0.45 | 0.60 |
| CNN w/o var. cluster; no context (spat. context N/A) | 0.60 | 0.14 | 0.23 |
| CNN w/o var. cluster; no context; with self-attention (spat. context N/A) | 0.64 | 0.21 | 0.32 |
| DCRNN w/o var. cluster; freq. context (spat. context N/A) | 1.00 | 0.50 | 0.67 |
| STGCN w/o var. cluster; freq. context (spat. context N/A) | 0.17 | 0.11 | 0.13 |
| GCN-LSTM w/o var. cluster; freq. context (spat. context N/A) | 1.00 | 0.17 | 0.29 |
| EEG and ICP | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 0.97 | 0.13 | 0.23 |
| LSTM w/o var. cluster; no context | 0.83 | 0.13 | 0.22 |
| LSTM w/o var. cluster; no context; with self-attention | 0.70 | 0.14 | 0.23 |
| CNN w/o var. cluster; no context | 0.67 | 0.08 | 0.14 |
| CNN w/o var. cluster; no context; with self-attention | 0.40 | 0.13 | 0.20 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.85 | 0.02 | 0.04 |
| STGCN w/o var. cluster; freq. & spat. context | 0.17 | 0.11 | 0.13 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.83 | 0.11 | 0.19 |
| EEG, ICP and ECG | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 1.00 | 0.17 | 0.29 |
| LSTM w/o var. cluster; no context | 0.80 | 0.17 | 0.28 |
| LSTM w/o var. cluster; no context; with self-attention | 0.40 | 0.18 | 0.25 |
| CNN w/o var. cluster; no context | 0.65 | 0.05 | 0.09 |
| CNN w/o var. cluster; no context; with self-attention | 0.40 | 0.19 | 0.26 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.93 | 0.02 | 0.04 |
| STGCN w/o var. cluster; freq. & spat. context | 0.17 | 0.04 | 0.06 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.17 | 0.20 | 0.18 |
| EEG, ICP, ECG and ABP | | | |
| LSTM-MMA w/o var. cluster; freq. & spat. context | 1.00 | 0.22 | 0.36 |
| LSTM w/o var. cluster; no context | 0.82 | 0.11 | 0.19 |
| LSTM w/o var. cluster; no context; with self-attention | 0.40 | 0.25 | 0.31 |
| CNN w/o var. cluster; no context | 0.78 | 0.04 | 0.08 |
| CNN w/o var. cluster; no context; with self-attention | 0.60 | 0.18 | 0.28 |
| DCRNN w/o var. cluster; freq. & spat. context | 0.88 | 0.08 | 0.15 |
| STGCN w/o var. cluster; freq. & spat. context | 0.17 | 0.03 | 0.05 |
| GCN-LSTM w/o var. cluster; freq. & spat. context | 0.17 | 0.14 | 0.15 |

Table 5 Comparison of LSTM-MMA against baselines for direct learning freq. & spat. context using EEG, ICP, ECG and ABP data (7 % rare event) – PCA reduction applied by default on RMT attention (the higher, the better)– $p_{max} = 5.1m$, $p_{min} = 4.4m$.

Competitor #2 (LSTM with Self-Attention). For the second baseline, we provide four segmented (one for EEG, ICP, ECG and ABP) LSTM without context, but we add attention using input data instead of RMT features as the query vector. LSTM output is used as the key and value vector thereby condensing attention into a single layer instead of the dual layer attention in LSTM-MMA. This baseline is a form of self-attention with dot product between (weighted) input data by (weighted) LSTM output followed by softmax operation to get attention vector which is applied on the (weighted) LSTM output. This baseline is different from the scaled dot-product self-attention used in Transformers with multiple attention heads [52].

Competitor #3 (Vanilla CNN). For the third baseline, we have CNN with four segmented convolution layers (one for EEG, ICP, ECG and ABP followed by a Batch Normalization layer); each convolution layer has 3 * 3 kernels,

output filters 20 and activation ReLU with zero padding (padded evenly such that output has the same dimension as the input). Vanilla CNN baseline has no context or MMA (Metadata Supported Multi-variate Attention).

Competitor #4 (CNN with Self-Attention). The CNN baseline without context is given self-attention in a similar manner as that of LSTM in the second baseline using input data instead of RMT features as the query vector and is considered as the fourth baseline. CNN output is the key and value vector here.

Competitor #5 (DCRNN). We consider Diffusion Convolution RNN (DCRNN [14]) as the fifth baseline against LSTM-MMA which takes into account both the spatial and temporal contexts. DCRNN is a graph convolution approach [14] that we have adapted for the seizure dataset with the same freq. & spat. context metadata matrix to serve as adjacency matrix, but it does not have the ability to rely on metadata supported, multi-variate attention (MMA) as proposed in this paper. The DCRNN model [14] is trained with a Gated Recurrent Unit (GRU). Number of diffusion hops for DCRNN is kept 1 for all datasets.

Competitor #6 (STGCN).

STGCN [58] is the sixth baseline against LSTM-MMA which takes into account both the spatial and temporal contexts. STGCN [58] is another graph convolution approach that we have adapted with the freq. & spat. context. The architecture of STGCN [58] consists of a graph convolutional layer using an adjacency matrix, temporal gated 1D convolutional layers, and fully connected layers. The graph convolutional layers operate on the spatial dimension of the EEG graph, allowing the model to aggregate information from neighboring spatial sensors and capture spatial dependencies. The temporal 1D convolutional layer for EEG operate on the output of graph convolution; each convolution layer has $3 * 3$ kernels, output filters 20 and activation ReLU with causal padding (padded the layer's input with zeros in the front to enable prediction of the values of early time steps), enabling the model to capture the temporal patterns and trends in the data. Above temporal convolution layer for EEG is multiplied element wise (Hadamard product) with another 1D temporal convolutional layer with a different activation; each convolution layer has $3 * 3$ kernels, output filters 20 and sigmoid activation with causal padding. Together these two temporal convolution layers form a gated convolution layer [58]. ICP, ECG and ABP also follow the same process but they have only gated temporal convolution layers as they do not have spatial context. The gated convolution layers are followed by a fully connected layer.

Competitor #7 (GCN-LSTM).

GCN-LSTM is the seventh baseline against LSTM-MMA which takes into account both the spatial and temporal contexts. The architecture of GCN-LSTM is adapted similar to STGCN [58] and consists of a graph convolutional layer using an adjacency matrix, temporal 1D convolutional layers, and fully connected layers. The graph convolutional layers operate on the spatial dimension of the EEG graph as in the case of STGCN [58]. The temporal 1D

convolutional layer for EEG also operate on the output of graph convolution; each convolution layer has $3 * 3$ kernels, output filters 20 and activation ReLU with causal padding as in the case of STGCN [58]. But in the GCN-LSTM model, after the first temporal convolution layer there is a LSTM layer (instead of another temporal convolution layer) with `return_sequence=True` parameter turned on. The output of convolution is fed to a LSTM layer for EEG inspired by [59]. ICP, ECG and ABP also follow the same process but they have only temporal convolution layers as they do not have spatial context.

DCRNN [14], STGCN [58] and GCN-LSTM [59] takes in same adjacency matrix for seizure dataset as in LSTM-MMA to find the neighbors of a EEG sensor. For the COVID, traffic and bitcoin datasets also, same spatial context metadata matrix is used in LSTM-MMA and baselines with context like DCRNN [14], STGCN [58] and GCN-LSTM [59]. Other baselines like attentioned and non-attentioned LSTM and CNN have no context at all. For all baselines with and without context, the timeseries data has segments of common sequence length for data preprocessing namely 500 (default unless specified) for the seizure dataset, 20 for the COVID dataset, 4 for the traffic dataset and 30 for the bitcoin dataset as is in the case of LSTM-MMA. The timeseries data sets are partitioned into a training set, validation set, and test set, with the same split ratio for all baselines as in LSTM-MMA.

We train baselines with the same hyperparameters as LSTM-MMA, given in Table 2, namely batch size (batch size=60), number of epochs (unless specified), and optimizer (Adam optimizer). A lower batch size of 30 is observed to be better for DCRNN as batch size of 60 sometimes caused memory issues for the seizure dataset. We report results using the accuracy measures defined in Sections 3.1.1, 3.1.2, 3.1.3 and 3.1.4 for all baselines.

3.3 Results

In this subsection, we present the results for seizure onset prediction, COVID, traffic, bitcoin forecasting tasks as detailed earlier. In the default experiments, we considered the version of the forecasting algorithm without variate clustering as we aim to observe the impact of the metadata supported MMA attention mechanism, without additional optimizations (such as variate clustering). We also present a separate ablation study which illustrates that variate clustering is an effective optimization technique (especially in the seizure detection data where the number of variates is very large).

3.3.1 Seizure Onset Prediction Task

Seizure Onset Prediction Task results are reported from Tables 3 through 12. Each table from Table 3 through 8 reports results on a single patient for direct learning. Tables 9 through 12 refer to transfer learning scenarios from a donor/provider patient to a test patient.

Comparison against the Baselines

As we see in Tables 3 through 5, the LSTM-MMA model is able to provide significantly better overall accuracy, when compared against the baselines.

- As mentioned in Section 3.1.1 there are two types of patients - patients with patterns of seizure and patients with one or more clusters of seizure events.
- For "ping-pong" seizure patient, results are reported in Table 3. When analyzing the EEG data for the "ping-pong" seizure, where the spatial context is important, DCRNN, STGCN and GCN-LSTM models are not able to provide good results – the proposed metadata supported multi-variate attention, however, enables the LSTM-MMA model to achieve relatively higher accuracy, especially high precision, for this scenario that requires spatial context. LSTM-MMA with *frequency and spatial context* is observed to be the best model for patients with patterns of seizure, best results are shown in bold. Self-attentioned CNN and LSTM models and GCN-LSTM have high recall for EEG but precision is lower than LSTM-MMA. The precision is low for DCRNN model for all signals whereas both the recall and precision are lower for STGCN model when compared to LSTM-MMA model for all signals.
- In contrast, in the case of "single long seizure" cluster patient, it is especially difficult to identify the single key event using spatial context with EEG signal and leverage it during model training, when compared to other patients having multiple seizures, as observed in Table 4. However for the ICP signal which has no spat. context, LSTM with self attention and DCRNN has similar (perfect) results similar to LSTM-MMA for patients with single seizure cluster as seen in Table 4. CNN model with self attention gives near perfect precision and recall for the ICP signal in Table 4. CNN model without context, GCN-LSTM and STGCN also has a high recall and precision with the ICP signal.
- In Table 5 in the case of "multiple seizure" cluster patient, the accuracy for EEG is better with spatial context when compared to "single long seizure" event patient, as there are multiple - three - key events. However ICP is the strong predictor here also for almost all models. This is firstly because the ICP signal does not have spatial context and because, unlike the ping-pong seizure, the seizure events are more homogeneous in these two cases in Tables 4 and 5 for patients with single and multiple seizure clusters.

Ablation Study – Model Direct Learning

In Tables 6 through 8, we analyze the impact of various components of LSTM-MMA through an ablation study:

- Table 6 presents results for the "ping-pong" seizure case which have multiple (seven) seizure onsets. As we see here, the best accuracies are obtained for all considered multi-modal scenarios when leveraging the spatial context of EEG, along with the underlying frequency context. Both recall and precision improved significantly when spatial context of EEG is considered for this usecase.

| Patient with "Ping-Pong" Seizure | | | |
|--|------------------|---------------------|------------------|
| Model | Mean CORecall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 0.52 | 0.85 | 0.65 |
| w/o var. cluster; freq. & spat. context | 0.83 | 0.91 | 0.87 |
| with var. cluster; freq. context | 0.67 | 0.70 | 0.68 |
| with var. cluster; freq. & spat. context | 0.75 | 0.70 | 0.72 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.50 | 0.67 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.50 | 0.67 |
| with var. cluster; freq. context | 1.00 | 0.80 | 0.89 |
| with var. cluster; freq. & spat. context | 1.00 | 0.80 | 0.89 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 0.52 | 0.71 | 0.60 |
| w/o var. cluster; freq. & spat. context | 0.87 | 1.00 | 0.93 |
| with var. cluster; freq. context | 0.75 | 0.59 | 0.66 |
| with var. cluster; freq. & spat. context | 0.90 | 0.56 | 0.69 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 0.54 | 0.53 | 0.53 |
| w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. context | 0.78 | 0.70 | 0.74 |
| with var. cluster; freq. & spat. context | 0.95 | 0.63 | 0.76 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 0.61 | 0.64 | 0.62 |
| w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. context | 0.73 | 0.65 | 0.69 |
| with var. cluster; freq. & spat. context | 0.90 | 0.67 | 0.77 |

Table 6 Ablation study using LSTM-MMA model for direct learning using EEG, ICP, ECG and ABP data (5.7% rare events, optimum $k = 50$ clusters) – PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

| Patient with Single Long Seizure | | | |
|--|------------------|---------------------|------------------|
| Model | Mean CORecall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 0.80 | 0.02 | 0.04 |
| w/o var. cluster; freq. & spat. context | 0.92 | 0.06 | 0.11 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.87 | 0.93 |
| w/o var. cluster; freq. & spat. context | 0.99 | 0.87 | 0.93 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 1.00 | 0.88 | 0.94 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.90 | 0.95 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.90 | 0.95 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.92 | 0.96 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |

Table 7 Ablation study using LSTM-MMA model for direct learning using EEG, ICP, ECG and ABP data (7% rare events, optimum $k = 100$ clusters) – PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

- As we see in Table 7, for the "single long seizure" case, the ICP signal provides the best (in fact perfect) accuracies. Even though the spatial context of EEG generally helps, the precision with the EEG signal is overall lower than that with the ICP. This is because, unlike the other patients (such as the ping-pong seizure patient), this patient has only a single onset that makes it difficult to discover the spatial context.

| Patient with Multiple Seizures | | | |
|--|------------------|---------------------|------------------|
| Model | Mean CORecall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 1.00 | 0.18 | 0.31 |
| w/o var. cluster; freq. & spat. context | 0.92 | 0.23 | 0.37 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.50 | 0.67 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.50 | 0.67 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 0.70 | 0.12 | 0.20 |
| w/o var. cluster; freq. & spat. context | 0.97 | 0.13 | 0.23 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 1.00 | 0.10 | 0.18 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.17 | 0.29 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.09 | 0.17 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.22 | 0.36 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |

Table 8 Ablation study using LSTM-MMA model for direct learning using EEG, ICP, ECG and ABP data (7% rare events, optimum $k = 50$ clusters) – PCA reduction applied by default on RMT attention (the higher, the better) – $pmax = 5.1m$, $pmin = 4.4m$.

- ICP is also the best signal for the "multiple seizures" cases which have multiple seizure onsets (Table 8). In this case, we see that spatial context underlying the EEG signal does help improve the accuracies for multi-modal scenarios involving EEG data streams.
- In general, the results with spatial context for EEG is observed to be best for the "ping-pong seizure" patient where spatial context is important as noted in Table 6. On the other hand, adaptive variate clustering done on EEG data streams gives perfect accuracy for patients with "single long seizure" cluster event and "multiple seizure" cluster events as observed in Tables 7 and 8 for both frequency and frequency and spatial contexts.

Ablation Study – Model Transfer Learning

We next investigate the impact of transferring models learned from one patient to another. In particular, we use the patient with "ping-pong" seizures, which show highest pattern diversity, as the donor and apply the learned model to other patients, with single, multiple, and lateral seizures.

- In Tables 9 to 12, we see that the model transfer is highly effective and that the spatial context captured by the metadata supported multi-variate attention (MMA) mechanism, proposed in this paper, generally helps with effective transfer of knowledge from one model to the other.
- ICP is the best signal in Table 9 on transfer learning to patient with "single long seizure" event as is the case in Table 7 on direct learning. Though spatial context helps to improve results for EEG alone scenario on transfer learning, ICP signal wins hands down as the best predictor for this patient with perfect accuracy for both frequency and frequency and spatial context.

| Patient with Single Long Seizure – Model Transferred from "Ping-Pong" Patient | | | |
|---|------------------|---------------------|------------------|
| Model | Mean CORecall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 0.88 | 0.18 | 0.30 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.21 | 0.35 |
| with var. cluster; freq. context | 1.00 | 0.58 | 0.73 |
| with var. cluster; freq. & spat. context | 1.00 | 0.78 | 0.88 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 0.98 | 0.13 | 0.23 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.13 | 0.23 |
| with var. cluster; freq. context | 1.00 | 0.85 | 0.92 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 0.90 | 0.07 | 0.13 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.05 | 0.10 |
| with var. cluster; freq. context | 1.00 | 0.93 | 0.96 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.08 | 0.15 |
| w/o var. cluster; freq. & spat. context | 0.96 | 0.05 | 0.10 |
| with var. cluster; freq. context | 1.00 | 0.93 | 0.96 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |

Table 9 Ablation study using LSTM-MMA model for transfer learning using EEG, ICP, ECG and ABP data (7% rare events, optimum $k = 200$ clusters) from "Ping-Pong" seizure patient – PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

- The patient with "multiple seizure" cluster events - three events - in Table 10 is noted to do better with transfer learning from ping-pong seizure patient using spatial context for all signals than by direct learning from this patient in Table 8.
- Another patient with "multiple seizure" cluster events - four events - and 0.8% anomaly in Table 11 also does well with transfer learning from ping-pong seizure patient using spatial context.
- Finally in Table 12 we transfer from "ping-pong" seizure patient to a patient with lateral seizures - having 7 events - and 0.5% anomaly. For this patient also transfer learning using spatial context of EEG works well.

Summary of the Seizure Prediction Experiments

The seizure prediction experiments reported so far has shown that the proposed LSTM-MMA approach with metadata-supported multi-variate attention provides significant gains in prediction accuracy against competitors. The ablation studies have further illustrated the effectiveness of the MMA approach in leveraging the frequency and spatial contexts provided by the metadata associated with the multivariate time series.

3.3.2 COVID Prediction Task

In this section, to illustrate the generalizability of the proposed techniques, we apply the proposed LSTM-MMA architecture to a different prediction problem. Table 13 shows prediction accuracies for LSTM-MMA and for baselines, with LSTM-MMA, DCRNN, STGCN and GCN-LSTM leveraging the spatial context

| Patient with Multiple Seizures – Model Transferred from "Ping-Pong" Patient | | | |
|---|------------------|---------------------|------------------|
| Model | Mean CORecall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 0.92 | 0.26 | 0.41 |
| w/o var. cluster; freq. & spat. context | 0.95 | 0.70 | 0.81 |
| with var. cluster; freq. context | 1.00 | 0.48 | 0.65 |
| with var. cluster; freq. & spat. context | 1.00 | 0.85 | 0.92 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.57 | 0.73 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.57 | 0.73 |
| with var. cluster; freq. context | 1.00 | 0.85 | 0.92 |
| with var. cluster; freq. & spat. context | 1.00 | 0.85 | 0.92 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.38 | 0.55 |
| w/o var. cluster; freq. & spat. context | 0.88 | 0.57 | 0.69 |
| with var. cluster; freq. context | 0.90 | 0.55 | 0.68 |
| with var. cluster; freq. & spat. context | 1.00 | 0.55 | 0.71 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 0.97 | 0.26 | 0.41 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.41 | 0.58 |
| with var. cluster; freq. context | 0.90 | 0.47 | 0.62 |
| with var. cluster; freq. & spat. context | 1.00 | 0.48 | 0.65 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.30 | 0.46 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.34 | 0.51 |
| with var. cluster; freq. context | 0.90 | 0.39 | 0.54 |
| with var. cluster; freq. & spat. context | 1.00 | 0.45 | 0.62 |

Table 10 Ablation study using LSTM-MMA model for transfer learning using EEG, ICP, ECG and ABP data (7% rare events, optimum $k = 200$ clusters) from "Ping-Pong" seizure patient– PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

in predicting rate of change of log of cases and rate of change of log of deaths in US States. These results do not consider variate clustering. As can be observed in Table 13, the RMSE and MSE are generally lower when we complement LSTM-MMA with MMA, which takes into account spatial context, and L1 regularization. LSTM-MMA without spatial context (implicit context with ones along the diagonal of the metadata matrix) with and without L1 regularization also have very low MSE and MAE values. STGCN, is in the third place, has significantly lower RMSE and MSE when compared to other baselines for predicting COVID rate of change of log of cases and deaths. DCRNN comes at the next place in predicting rate of change of log of cases and deaths. LSTM without context with self attention and GCN-LSTM have a similar MSE as that of DCRNN in predicting rate of change of log of deaths.

3.3.3 Traffic Forecasting Task

Results for traffic forecasting task are observed in Table 14. The RMSE and MSE are lower when we complement LSTM-MMA with MMA, which takes into account spatial context, and L1 regularization. These results do not consider variate clustering. LSTM-MMA without spatial context with L1 (implicit context with ones along the diagonal of the metadata matrix), LSTM-MMA with and without spatial context (implicit and explicit spatial context) without L1 regularization, CNN based baselines like CNN without any context but with and without self-attention and STGCN also fares well, have significantly lower

| Patient with Multiple Seizures (0.8% anomaly) – Model Transferred from "Ping-Pong" Patient | | | |
|--|-----------------|---------------------|------------------|
| Model | Mean CORcall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 0.90 | 0.62 | 0.73 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.71 | 0.83 |
| with var. cluster; freq. context | 1.00 | 0.80 | 0.89 |
| with var. cluster; freq. & spat. context | 1.00 | 0.85 | 0.92 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| w/o var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. context | 1.00 | 1.00 | 1.00 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.84 | 0.91 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.92 | 0.96 |
| with var. cluster; freq. context | 1.00 | 0.85 | 0.92 |
| with var. cluster; freq. & spat. context | 1.00 | 0.95 | 0.97 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 1.00 | 0.58 | 0.73 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.78 | 0.88 |
| with var. cluster; freq. context | 1.00 | 0.76 | 0.86 |
| with var. cluster; freq. & spat. context | 1.00 | 0.87 | 0.93 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 0.93 | 0.74 | 0.82 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.70 | 0.82 |
| with var. cluster; freq. context | 1.00 | 0.66 | 0.80 |
| with var. cluster; freq. & spat. context | 1.00 | 0.71 | 0.83 |

Table 11 Ablation study using LSTM-MMA model for transfer learning using EEG, ICP, ECG and ABP data (2% and lower rare events, optimum $k = 250$ clusters) from "Ping-Pong" seizure patient– PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

MSE when compared to LSTM based baselines like GCN-LSTM, LSTM without context without self-attention and GRU based baseline like DCRNN for forecasting traffic flow.

3.3.4 Bitcoin Forecasting Task

Results for bitcoin forecasting task are observed in Table 15. These results do not consider variate clustering. As can be observed, the RMSE and MSE are lowest, when we complement LSTM-MMA with MMA, which takes into account metadata supported correlation context and L1 regularization, closely followed by LSTM-MMA with context but without L1 regularization. LSTM-MMA without context (implicit context with ones along the diagonal of the metadata matrix) with and without L1 regularization, LSTM without any context with self-attention also have lower RMSE, MSE, and MAE when compared to CNN without context, GCN-LSTM, DCRNN and STGCN baselines for the bitcoin dataset.

4 Conclusions

Post traumatic seizure prediction tasks are hampered by the rareness of such events. Arguing that multi-variate multi-modal time series carry robust localized temporal and spatial features that could help identify these rare seizure events, we proposed a metadata supported multi-variate attention (or MMA) technique, which leverages robust multi-variate temporal and spatial features, and presented an LSTM-based architecture to predict onset of seizure events.

| Patient with Lateral Seizures(0.5% anomaly)– Model Transferred from "Ping-Pong" Patient | | | |
|---|-----------------|---------------------|------------------|
| Model | Mean CORcall | Mean COPrecision | Mean F1 Score |
| EEG | | | |
| w/o var. cluster; freq. context | 1.00 | 0.42 | 0.59 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.75 | 0.86 |
| with var. cluster; freq. context | 0.80 | 0.24 | 0.37 |
| with var. cluster; freq. & spat. context | 1.00 | 1.00 | 1.00 |
| ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.33 | 0.50 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.33 | 0.50 |
| with var. cluster; freq. context | 1.00 | 0.39 | 0.56 |
| with var. cluster; freq. & spat. context | 1.00 | 0.39 | 0.56 |
| EEG and ICP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.34 | 0.51 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.36 | 0.53 |
| with var. cluster; freq. context | 0.80 | 0.08 | 0.15 |
| with var. cluster; freq. & spat. context | 0.95 | 0.59 | 0.73 |
| EEG, ICP and ECG | | | |
| w/o var. cluster; freq. context | 1.00 | 0.34 | 0.51 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.37 | 0.54 |
| with var. cluster; freq. context | 0.80 | 0.08 | 0.15 |
| with var. cluster; freq. & spat. context | 0.98 | 0.75 | 0.85 |
| EEG, ICP, ECG and ABP | | | |
| w/o var. cluster; freq. context | 1.00 | 0.42 | 0.59 |
| w/o var. cluster; freq. & spat. context | 1.00 | 0.78 | 0.88 |
| with var. cluster; freq. context | 0.80 | 0.08 | 0.15 |
| with var. cluster; freq. & spat. context | 1.00 | 0.78 | 0.88 |

Table 12 Ablation study using LSTM-MMA model for transfer learning using EEG, ICP, ECG and ABP data (2% and lower rare events, optimum $k = 100$ clusters) from "Ping-Pong" seizure patient– PCA reduction applied by default on RMT attention (the higher, the better)– $pmax = 5.1m$, $pmin = 4.4m$.

Experiments on EEG, ICP, ECG and ABP data show that the proposed LSTM-MMA model is highly effective in improving model accuracy for rare event early prediction tasks. Generally we observed that the model learns robust features well when EEG and other modality data are smoothed with variate clustering and when given frequency and spatial context to the modalities. We also observed that while the EEG signal may sometimes be ineffective, there are secondary modalities like ICP that are strong predictors (with best F1 score) for predicting seizure onset events.

Additional experiments done on publicly available multi-variate COVID, traffic, bitcoin datasets show that LSTM-MMA model is effective on regression tasks as well.

5 Acknowledgments

Thanks to Dr Stephen Foldes and Austin Jacobson of Phoenix Children's, Phoenix, AZ for their support with the work. Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation.

6 Compliance with Ethical Standards

- *I. Ethical approval.* For data collection and analysis, the research is approved by the Phoenix Children's Institutional Review Board (IRB #19-284).

| Rate of change of log of cases | | | |
|--|---------------|---------------|---------------|
| Model | RMSE | MSE | MAE |
| LSTM-MMA w/o spat. context | 0.0102 | 0.0001 | 0.0044 |
| LSTM-MMA w/o spat. context; with L1 | 0.0104 | 0.0001 | 0.0044 |
| LSTM-MMA with spat. context | 0.0103 | 0.0001 | 0.0082 |
| LSTM-MMA with spat. context; with L1 | 0.0101 | 0.0001 | 0.0078 |
| LSTM w/o var. cluster; no context | 0.0374 | 0.0014 | 0.0320 |
| LSTM w/o var. cluster; no context; with self-attention | 0.0361 | 0.0013 | 0.0344 |
| CNN w/o var. cluster; no context | 0.1024 | 0.0105 | 0.0794 |
| CNN w/o var. cluster; no context; with self-attention | 0.0316 | 0.0010 | 0.0291 |
| DCRNN with spat. context | 0.0192 | 0.0004 | 0.0138 |
| STGCN with spat. context | 0.0141 | 0.0002 | 0.0096 |
| GCN-LSTM with spat. context | 0.0436 | 0.0019 | 0.0394 |

| Rate of change of log of deaths | | | |
|--|---------------|---------------|---------------|
| Model | RMSE | MSE | MAE |
| LSTM-MMA w/o spat. context | 0.0141 | 0.0002 | 0.0087 |
| LSTM-MMA w/o spat. context; with L1 | 0.0141 | 0.0002 | 0.0087 |
| LSTM-MMA with spat. context | 0.0120 | 0.0001 | 0.0073 |
| LSTM-MMA with spat. context; with L1 | 0.0119 | 0.0001 | 0.0073 |
| LSTM w/o var. cluster; no context | 0.0308 | 0.0009 | 0.0263 |
| LSTM w/o var. cluster; no context; with self-attention | 0.0265 | 0.0007 | 0.0238 |
| CNN w/o var. cluster; no context | 0.2262 | 0.0512 | 0.1813 |
| CNN w/o var. cluster; no context; with self-attention | 0.0849 | 0.0072 | 0.0720 |
| DCRNN with spat. context | 0.0258 | 0.0007 | 0.0178 |
| STGCN with spat. context | 0.0141 | 0.0002 | 0.0091 |
| GCN-LSTM with spat. context | 0.0265 | 0.0007 | 0.0244 |

Table 13 Comparison of the LSTM-MMA model against baselines for COVID prediction – PCA reduction applied by default on MMA (the lower, the better) – for next $p = 1$ day.

| Traffic travel time forecasting | | | |
|---------------------------------------|-------------|--------------|-------------|
| Model | RMSE | MSE | MAE |
| LSTM-MMA w/o spat. context | 4.77 | 22.76 | 3.33 |
| LSTM-MMA w/o spat. context; with L1 | 4.75 | 22.52 | 3.34 |
| LSTM-MMA with spat. context | 4.75 | 22.61 | 3.31 |
| LSTM-MMA with spat. context; with L1 | 4.74 | 22.48 | 3.28 |
| LSTM w/o context | 7.39 | 54.65 | 5.61 |
| LSTM w/o context; with self-attention | 6.47 | 41.81 | 4.91 |
| CNN w/o context | 5.79 | 33.50 | 4.21 |
| CNN w/o context; with self-attention | 5.37 | 28.79 | 3.96 |
| DCRNN with spat. context | 6.69 | 44.76 | 5.09 |
| STGCN with spat. context | 5.74 | 32.91 | 4.07 |
| GCN-LSTM with spat. context | 13.79 | 190.21 | 12.33 |

Table 14 Comparison of the LSTM-MMA model against baselines for Traffic forecasting – PCA reduction applied by default on MMA (the lower, the better) – for next $p = 15$ minutes.

| Bitcoin price forecasting | | | |
|--------------------------------------|---------------|--------------------------------------|---------------|
| Model | RMSE | MSE | MAE |
| LSTM-MMA w/o context | 0.0013 | $1.6652 * 10^{-6}$ | 0.0011 |
| LSTM-MMA w/o context; with L1 | 0.0012 | $1.5550 * 10^{-6}$ | 0.0011 |
| LSTM-MMA with context | 0.0012 | $1.4469 * 10^{-6}$ | 0.0010 |
| LSTM-MMA with context; with L1 | 0.0012 | $1.3580 * 10^{-6}$ | 0.0010 |
| LSTM w/o context | 0.0013 | $1.7531 * 10^{-6}$ | 0.0012 |
| LSTM w/o context with self-attention | 0.0013 | $1.6288 * 10^{-6}$ | 0.0011 |
| CNN w/o context | 0.0013 | $1.7674 * 10^{-6}$ | 0.0012 |
| CNN w/o context with self-attention | 0.0013 | $1.7523 * 10^{-6}$ | 0.0012 |
| DCRNN with context | 0.0014 | $1.8660 * 10^{-6}$ | 0.0012 |
| STGCN with context | 0.0014 | $2.0544 * 10^{-6}$ | 0.0013 |
| GCN-LSTM with context | 0.0013 | $1.7690 * 10^{-6}$ | 0.0012 |

Table 15 Comparison of the LSTM-MMA model against baselines for Bitcoin forecasting – PCA reduction applied by default on MMA (the lower, the better) – from $p = 30$ days onwards.

- *II. Funding details* The work is primarily supported by Department Of Defense grant W81XWH-19-1-0514. Some aspects of the research is also supported by NSF grants #1909555, #2026860, #1827757.
- *III. Conflict of interest.* The authors declare that they have no conflict of interest.
- *IV. Informed consent.* For retrospective analysis as ours, informed consent was not required by the the Phoenix Children’s Institutional Review Board.

References

- [1] Epilepsy foundation, michigan. <https://epilepsymichigan.org/page.php?id=358>
- [2] Ding, K., Gupta, P.K., Diaz-Arrastia, R.: Epilepsy after traumatic brain injury. *Translational research in traumatic brain injury* (2016)
- [3] Klem, G.H., Lüders, H.O., Jasper, H., Elger, C., *et al.*: The ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol* **52**(3), 3–6 (1999)
- [4] Tong, Y.: Seizures caused by pyridoxine (vitamin b6) deficiency in adults: A case report and literature review. *Intractable & rare diseases research* **3**(2), 52–56 (2014)
- [5] Wang, H.-S., Kuo, M.: Vitamin b⁶ related epilepsy during childhood. *Chang Gung medical journal* **30**(5), 396 (2007)
- [6] Kirik, S., Çatak, Z.: Vitamin b12 deficiency observed in children with first afebrile seizures. *Cureus* **13**(3) (2021)
- [7] Natu, M., Bachute, M., Gite, S., Kotecha, K., Vidyarthi, A.: Review on epileptic seizure prediction: machine learning and deep learning approaches. *Computational and Mathematical Methods in Medicine* **2022** (2022)
- [8] Fan, M., Chou, C.-A.: Detecting abnormal pattern of epileptic seizures via temporal synchronization of eeg signals. *IEEE Transactions on Biomedical Engineering* **66**(3), 601–608 (2018)
- [9] Shahsavari, S., McKelvey, T., Ritzen, C.E., Rydenhag, B.: Cerebrovascular mechanical properties and slow waves of intracranial pressure in tbi patients. *IEEE Transactions on Biomedical Engineering* **58**(7), 2072–2082 (2011)
- [10] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T.-Y.: Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823* (2020)

- [11] Ravindranath, M., Candan, K.S., Sapino, M.L.: M2nn: Rare event inference through multi-variate multi-scale attention. In: 2020 IEEE International Conference on Smart Data Services (SMDS), pp. 53–62 (2020). IEEE
- [12] Liu, S., Poccia, S.R., Candan, K.S., Sapino, M.L., Wang, X.: Robust multi-variate temporal features of multi-variate time series. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **14**(1), 7 (2018)
- [13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011)
- [14] Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017)
- [15] Valderrama, M., Nikolopoulos, S., Adam, C., Navarro, V., Le Van Quyen, M.: Patient-specific seizure prediction using a multi-feature and multi-modal eeg-ecg classification. In: XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010, pp. 77–80 (2010). Springer
- [16] Li, G., Lee, C.H., Jung, J.J., Youn, Y.C., Camacho, D.: Deep learning for eeg data analytics: A survey. *Concurrency and Computation: Practice and Experience* **32**(18), 5199 (2020)
- [17] Usman, S.M., Khalid, S., Aslam, M.H.: Epileptic seizures prediction using deep learning techniques. *Ieee Access* **8**, 39998–40007 (2020)
- [18] Park, Y., Luo, L., Parhi, K.K., Netoff, T.: Seizure prediction with spectral power of eeg using cost-sensitive support vector machines. *Epilepsia* **52**(10), 1761–1770 (2011)
- [19] Ramgopal, S., Thome-Souza, S., Jackson, M., Kadish, N.E., Fernández, I.S., Klehm, J., Bosl, W., Reinsberger, C., Schachter, S., Loddenkemper, T.: Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & behavior* **37**, 291–307 (2014)
- [20] Usman, S.M., Usman, M., Fong, S.: Epileptic seizures prediction using machine learning methods. *Computational and mathematical methods in medicine* **2017** (2017)
- [21] Alotaiby, T.N., Alshebeili, S.A., Alotaibi, F.M., Alrshoud, S.R.: Epileptic seizure prediction using csp and lda for scalp eeg signals. *Computational*

- intelligence and neuroscience **2017** (2017)
- [22] Islam, M., Tanaka, T., Iimura, Y., Mitsuhashi, T., Sugano, H., Wang, D., Molla, M., Islam, K., *et al.*: Statistical features in high-frequency bands of interictal i EEG work efficiently in identifying the seizure onset zone in patients with focal epilepsy. *Entropy* **22**(12), 1415 (2020)
- [23] Patel, K.S., Zhao, M., Ma, H., Schwartz, T.H.: Imaging preictal hemodynamic changes in neocortical epilepsy. *Neurosurgical focus* **34**(4), 10 (2013)
- [24] Storti, S.F., Del Felice, A., Formaggio, E., Boscolo Galazzo, I., Bongiovanni, L.G., Cerini, R., Fiaschi, A., Manganotti, P.: Spatial and temporal eeg-fMRI changes during preictal and postictal phases in a patient with posttraumatic epilepsy. *Clinical EEG and neuroscience* **46**(3), 247–252 (2015)
- [25] Ozcan, A.R., Erturk, S.: Seizure prediction in scalp eeg using 3d convolutional neural networks with an image-based approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27**(11), 2284–2293 (2019)
- [26] Liu, L., Chen, W., Cao, G.: Prediction of neonatal amplitude-integrated eeg based on lstm method. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 497–500 (2016). IEEE
- [27] Tsiouris, K.M., Pezoulas, V.C., Zervakis, M., Konitsiotis, S., Koutsouris, D.D., Fotiadis, D.I.: A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals. *Computers in biology and medicine* **99**, 24–37 (2018)
- [28] Singh, K., Malhotra, J.: Two-layer lstm network-based prediction of epileptic seizures using eeg spectral features. *Complex & Intelligent Systems*, 1–14 (2022)
- [29] Lun, X., Yu, Z., Chen, T., Wang, F., Hou, Y.: A simplified cnn classification method for mi-eeg via the electrode pairs signals. *Frontiers in Human Neuroscience* **14**, 338 (2020)
- [30] Jiang, Y., Wu, D., Deng, Z., Qian, P., Wang, J., Wang, G., Chung, F.-L., Choi, K.-S., Wang, S.: Seizure classification from eeg signals using transfer learning, semi-supervised learning and tsf fuzzy system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **25**(12), 2270–2284 (2017)
- [31] Pinto, M., Coelho, T., Leal, A., Lopes, F., Dourado, A., Martins, P., Teixeira, C.: Interpretable eeg seizure prediction using a multiobjective

- evolutionary algorithm. *Scientific reports* **12**(1), 1–15 (2022)
- [32] Stacey, W.C.: Seizure prediction is possible—now let’s make it practical. *EBioMedicine* **27**, 3–4 (2018)
- [33] Dumanis, S.B., French, J.A., Bernard, C., Worrell, G.A., Fureman, B.E.: Seizure forecasting from idea to reality. outcomes of the my seizure gauge epilepsy innovation institute workshop. *eneuro* **4**(6) (2017)
- [34] Hain, D., Jurowetzki, R.: Introduction to rare-event predictive modeling for inferential statisticians—a hands-on application in the prediction of breakthrough patents. *arXiv preprint arXiv:2003.13441* (2020)
- [35] Cheon, S.-P., Kim, S., Lee, S.-Y., Lee, C.-B.: Bayesian networks based rare event prediction with sensor data. *Knowledge-Based Systems* **22**(5), 336–343 (2009)
- [36] Li, J., Liu, L.-s., Fong, S., Wong, R.K., Mohammed, S., Fiaidhi, J., Sung, Y., Wong, K.K.: Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PloS one* **12**(7), 0180830 (2017)
- [37] Li, J., Fong, S., Mohammed, S., Fiaidhi, J., Chen, Q., Tan, Z.: Solving the under-fitting problem for decision tree algorithms by incremental swarm optimization in rare-event healthcare classification. *Journal of Medical Imaging and Health Informatics* **6**(4), 1102–1110 (2016)
- [38] Li, J., Fong, S., Hu, S., Chu, V.W., Wong, R.K., Mohammed, S., Dey, N.: Rare event prediction using similarity majority under-sampling technique. In: *International Conference on Soft Computing in Data Science*, pp. 23–39 (2017). Springer
- [39] Ando, S., Huang, C.Y.: Deep over-sampling framework for classifying imbalanced data. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 770–785 (2017). Springer
- [40] Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y.-G., Ding, K., Chen, Z.: Trainable undersampling for class-imbalance learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4707–4714 (2019)
- [41] Bao, F., Deng, Y., Kong, Y., Ren, Z., Suo, J., Dai, Q.: Learning deep landmarks for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems* **31**(8), 2691–2704 (2019)
- [42] Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data.

- IEEE transactions on neural networks and learning systems **29**(8), 3573–3587 (2017)
- [43] Kim, J., Jeong, J., Shin, J.: M2m: Imbalanced classification via major-to-minor translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13896–13905 (2020)
- [44] Ranjan, C., Reddy, M., Mustonen, M., Paynabar, K., Pourak, K.: Dataset: rare event classification in multivariate time series. arXiv preprint arXiv:1809.10717 (2018)
- [45] Xu, D., Zhang, Z., Shi, J.: Training data selection by categorical variables for better rare event prediction in multiple products production line. *Electronics* **11**(7), 1056 (2022)
- [46] Ali, Ö.G., Artürk, U.: Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications* **41**(17), 7889–7903 (2014)
- [47] Xiu, Z., Tao, C., Gao, M., Davis, C., Goldstein, B.A., Henao, R.: Variational disentanglement for rare event modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 10469–10477 (2021)
- [48] Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Advances in Neural Information Processing Systems, pp. 1410–1418 (2009)
- [49] Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems, pp. 935–943 (2013)
- [50] Dong, Y., Jiang, X., Zhou, H., Lin, Y., Shi, Q.: Sr2cnn: Zero-shot learning for signal recognition. *IEEE Transactions on Signal Processing* **69**, 2316–2329 (2021)
- [51] Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733 (2016)
- [52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
- [53] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
- [54] Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches

- to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
- [55] Khattar, A., Quadri, S.: Camm: Cross-attention multimodal classification of disaster-related tweets. *IEEE Access* **10**, 92889–92902 (2022)
- [56] Garg, Y., Candan, K.S.: Xm2a: Multi-scale multi-head attention with cross-talk for multi-variate time series analysis. In: 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 151–157 (2021). IEEE
- [57] Garg, Y., Candan, K.S.: Sdma: Saliency-driven mutual cross attention for multi-variate time series. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7242–7249 (2021). IEEE
- [58] Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875 (2017)
- [59] Bogaerts, T., Masegosa, A.D., Angarita-Zapata, J.S., Onieva, E., Hellinckx, P.: A graph cnn-lstm neural network for short and long-term traffic forecasting based on trajectory data. *Transportation Research Part C: Emerging Technologies* **112**, 62–77 (2020)
- [60] Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288 (1996)
- [61] Freedman, D., Pisani, R., Purves, R.: *Statistics: Fourth international student edition*. WW Nort Co <https://www.Amaz.Com/Statistics-Fourth-Int-Stud-Free> Accessed **22** (2020)
- [62] The new york times coronavirus (covid-19) cases and deaths in the united states. <https://data.humdata.org/dataset/nyt-covid-19-data>
- [63] Population, population change, and estimated components of population change: April 1, 2010 to july 1, 2019 (nst-est2019-alldata). <https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-total.html>
- [64] England, H.: *Highways agency network journey time and traffic flow data*. Highways England: Guildford, UK (2018)
- [65] Graves, A., Graves, A.: Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37–45 (2012)